Measurement of Population Diversity

Ronald W. Morrison[1] and Kenneth A. De Jong[2]

[1] Mitretek Systems, Inc.
7525 Colshire Drive
McLean, VA 22102-7400
ronald.morrison@mitretek.org
[2] Department of Computer Science, George Mason University
Fairfax, VA 22030
kdejong@gmu.edu

Abstract. In evolutionary algorithms (EAs), the need to efficiently measure population diversity arises in a variety of contexts, including operator adaptation, algorithm stopping and re-starting criteria, and fitness sharing. In this paper we introduce a unified measure of population diversity and define its relationship to the most common phenotypic and genotypic diversity measures. We further demonstrate that this new measure provides a new and efficient method for computing population diversity, where the cost of computation increases linearly with population size.

1 Introduction

Population diversity is a key measurement in a variety of EA implementations. The question of when to stop the EA or when to re-start the EA is often based on a measure of population diversity. In fitness sharing algorithms, population diversity is used as a basis for distributing the fitness credit. The use of EAs for dynamic fitness landscapes requires measures for maintaining population diversity to ensure that the EA can detect and respond to the changes in the landscape.

Several methods for estimating population diversity have been used. They include diversity measures in both genotypic space and phenotypic space. In phenotypic space, several pair-wise and "column-based" measures (measuring the variation in values for each specific phenotypic feature) have been suggested (e.g., [1]). Genotypic measures are much more common. Principal genotypic measures are entropy (e.g., [2]), and, more commonly, pair-wise Hamming distance (e.g., [3]). Pair-wise Hamming distance H of P strings of length L is defined as:

$$H = \sum_{j=1}^{j=P-1} \sum_{j'=j+1}^{j'=P} \left(\sum_{i=1}^{i=L} |y_{ij} - y_{ij'}|\right) \tag{1}$$

where $y_{ij}, y_{ij'} \in \{0, 1\}$ and the generalized notation,

$$\sum_{k=1}^{k=M-1} \sum_{k'=k+1}^{k'=M} f(x_k, x_{k'}) \tag{2}$$

is the sum of the results of the application of $f(x_k, x_{k'})$ to all pair-wise combinations the members x_k and $x_{k'}$ of a given population of size M.

Historically, one of the major difficulties in the use of pair-wise population diversity measures is that the computation of the measure is quadratic with the size of the population P for pair-wise selection:

$$\binom{P}{2} = \frac{P^2 - P}{2}. \tag{3}$$

In this paper we introduce a unified measure of population diversity and define its relationship between the most common phenotypic and genotypic diversity measures. We further demonstrate that this new measure provides a new and efficient method for computing population diversity, where the cost of computation increases linearly with population size. Section 2 of the paper will provide background information; Section 3 will define the diversity measure; Section 4 will relate the new diversity measure to other diversity measure in genotypic space; Section 5 will discuss the the diversity measure's relationship to other phenotypic-space measures; and Section 6 provides the conclusions and discusses future work.

2 Background

2.1 Historical Measures of Population Diversity

The most commonly used measures of population diversity include pair-wise Hamming distance in genotypic space, and column-based pair-wise distance and column variance in phenotypic space. In real-number optimization problems, phenotypic space diversity measures are often preferred over binary encoded genotypic measures. This is because, when using genotypic measures, all bit-wise diversity is treated the same, but variations at the different bit positions can represent significantly different levels of phenotypic diversity. Figures 1 through 9 provide illustrations of the three common diversity measures, using a simple genetic algorithm (GA), a population of 20 on a 2-dimensional, multi-modal landscape similar to that described in [4]. Gray code was used for the binary representation for the GA. Figure 1 is the initial population distribution. Figures 2 through 4 show the convergence of the population at generations 5, 16 and 20 respectively.

Figure 5 shows the pair-wise Hamming distance at each generation. Figure 6 provides the sum of the pair-wise distances of each column, and Figure 7 provides the sum of the variances of each column.

As can be seen in Figure 3, the population has lost nearly all of its diversity by generation 16. The three diversity measures provide somewhat different views of this loss of diversity, with the column variances (Figure 7) most clearly indicating population convergence, while the low-order bit differences cause the genotypic space pair-wise Hamming distance measure (Figure 5) to indicate more diversity than is present in phenotypic space.

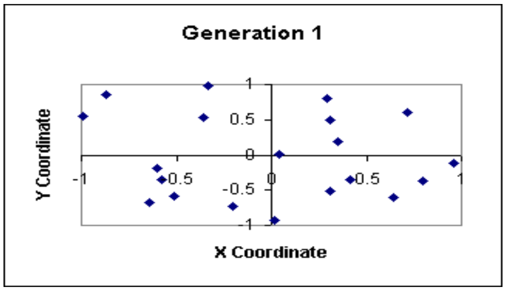

Fig. 1. Population at Generation=1

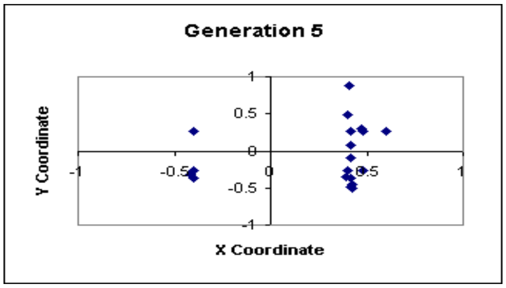

Fig. 2. Population at Generation=5

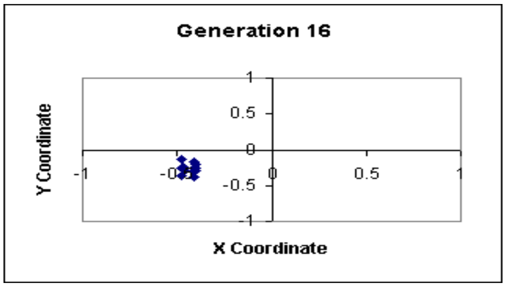

Fig. 3. Population at Generation=16

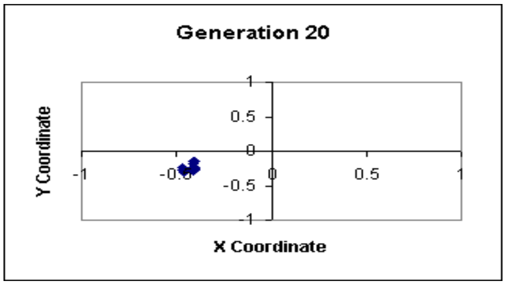

Fig. 4. Population at Generation=20

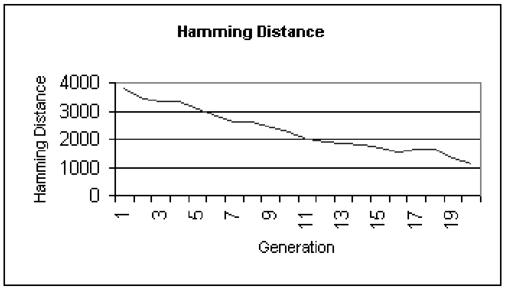

Fig. 5. Population Pair-wise Hamming Distance

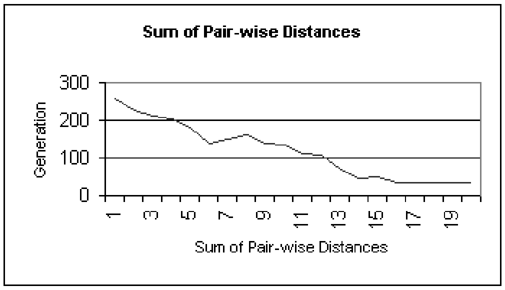

Fig. 6. Sum of Pair-wise Distances

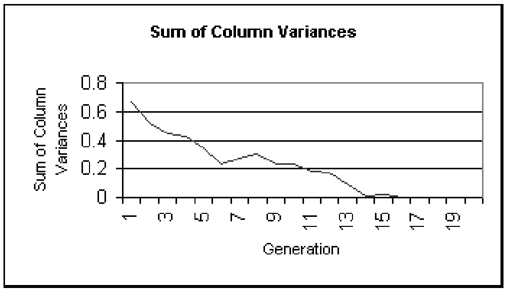

Fig. 7. Sum of Column Variances

2.2 Concept Review

The new population diversity measure to be presented herein is derived from some traditional engineering concepts that we have adapted to this problem. To facilitate the upcoming discussion, a brief review of these concepts is provided.

The first concept of interest is the centroid. The centroid of an object, also called the center of mass or center of gravity, is the point of balance for the entire object. The coordinates of the centroid are the coordinates of the midpoints of the mass distribution along each axis.

The second concept of interest is the moment of inertia. Moment of inertia is a term used in many engineering problems and calculations. Just as mass is the relationship between force and acceleration according to Newton's second law, moment of inertia is the relationship between torque and angular acceleration. The moment of inertia indicates how easily an object rotates about a point of rotation. In any object, the parts that are farthest from the axis of rotation contribute more to the moment of inertia than the parts that are closer to the axis. Conceptually, when the point of rotation is the centroid of an object, the moment of inertia is a measure of how far the mass of the object is distributed from the center of gravity of the object. The engineering moment of inertia for a point mass is defined as:

$$I = mr^2 \tag{4}$$

where: I is the usual symbol for moment of inertia, m is the mass, and r^2 is the square of the distance to the point of rotation.

3 A New Measure of Diversity

Our new measure of population diversity is based on extension of the concept of moment of inertia for measurement of mass distribution into arbitrarily high dimensionality spaces for the measurement of EA population diversity.

Extended into N-space, the coordinates of the centroid of P equally weighted points in N-space, $C = (c_1, c_2, c_3, \ldots c_N)$, are computed as follows:

$$c_i = \frac{\sum_{j=1}^{j=P} x_{ij}}{P} \tag{5}$$

where $x_{ij} \in \Re$ and c_i is the ith coordinate of the centroid.

Continuing with P equally-weighted points in N-space, we define the moment-of-inertia based measure of diversity of these points about their centroid is:

$$I = \sum_{i=1}^{i=N} \sum_{j=1}^{j=P} (x_{ij} - c_i)^2. \tag{6}$$

As will be shown in later sections, this measurement of population diversity is closely related to commonly used measures of both genotypic diversity and phenotypic diversity, providing a single diversity measurement method for use in both situations. The principal advantage of this measure of diversity is that, in comparison with traditional methods of computing pair-wise population diversity which are quadratic in population size, P, this method is linear in P. Specifically, for an N-dimensional problem with a population size of P, computation of the coordinates of the centroid requires N times P additions and N divisions. Computation of the moment of inertia around the centroid is then N times P subtractions plus N times P multiplications plus N times P additions. Total computational requirements for the centroid-based moment of inertia, therefore are $4(NP) + N$ calculations, making it a computationally efficient diversity measure.

4 Relationship to Diversity Measures in Genotypic Space

Genotypic diversity of EAs is most often measured using pair-wise Hamming distance, but the population diversity is much more efficiently computed using the new moment of inertia method.

When applying the moment of inertia calculation in the context of binary strings, each bit is assumed to be an independent "spatial" dimension. Under these circumstances, the coordinates of the centroid, $(c_1, c_2, c_3, \ldots, c_L)$, of P bit strings of length L are computed as:

$$c_i = \frac{\sum_{j=1}^{j=P} x_{ij}}{P} \tag{7}$$

and the moment of inertia about the centroid is:

$$I = \sum_{i=1}^{i=L} \sum_{j=1}^{j=P} (x_{ij} - c_i)^2. \tag{8}$$

It turns out that by transitioning from discrete mathematics to continuous mathematics, it can be shown that the moment of inertia as described in equation (8) is equal to the pair-wise Hamming distance divided by the population size.

Theorem 1. *For $y_{ij} \in \{0, 1\}$:*

$$\sum_{i=1}^{i=L} \sum_{j=1}^{j=P-1} \sum_{j'=j+1}^{j'=P} |y_{ij} - y_{ij'}| = P[\sum_{i=1}^{i=L} \sum_{j=1}^{j=P} (y_{ij} - c_i)^2] \tag{9}$$

where:

$$c_i = \frac{\sum_{j=1}^{j=P} x_{ij}}{P}.$$

Verbally: the pair-wise Hamming distance for P bit strings of length L is equal to the L-space moment of inertia of the population computed around the centroid of the population times the population size. In short, the pair-wise Hamming distance is the binary case of the centroid moment of inertia.

Proof. [1]
First we will examine the right hand side of the theorem:

$$P \sum_{i=1}^{L} \sum_{j=1}^{P} (y_{ij} - c_j)^2 = P \sum_{i=1}^{L} \sum_{j=1}^{P} (y_{ij} - \frac{\sum_{j=1}^{P} y_{ij}}{P})^2$$

$$= P \sum_{i=1}^{L} \sum_{j=1}^{P} (y_{ij}^2 - 2y_{ij} \frac{\sum_{j=1}^{P} y_{ij}}{P} + \frac{1}{P^2} (\sum_{j=1}^{P} y_{ij})^2)$$

[1] Proof based on suggestions by Chris Reedy, Mitretek Systems.

$$= P\sum_{i=1}^{L}[\sum_{j=1}^{P} y_{ij}^2 - \frac{2}{P}(\sum_{j}^{P} y_{ij})^2 + \frac{1}{P^2}\sum_{j=1}^{P}(\sum_{j=1}^{P} y_{ij})^2]$$

$$= P\sum_{i=1}^{L}[\sum_{j=1}^{P} y_{ij}^2 - \frac{2}{P}(\sum_{j=1}^{P} y_{ij})^2 + \frac{1}{P}(\sum_{j=1}^{P} y_{ij})^2]$$

$$= P\sum_{i=1}^{L}[\sum_{j=1}^{P} y_{ij}^2 - \frac{1}{P}(\sum_{j=1}^{P} y_{ij})^2] = P\sum_{i=1}^{L}\sum_{j=1}^{P} y_{ij}^2 - \sum_{i=1}^{L}(\sum_{j=1}^{P} y_{ij})^2. \tag{10}$$

To examine the left hand side of the theorem, let's first examine the properties of the quantity:

$$\sum_{i=1}^{L}\sum_{j=1}^{P}\sum_{j'=1}^{P}(y_{ij}-y_{ij'})^2 = \sum_{i=1}^{L}\sum_{j=1}^{P}\sum_{j'=1}^{P} y_{ij}^2 - 2\sum_{i=1}^{L}\sum_{j=1}^{P}\sum_{j'=1}^{P} y_{ij}y_{ij'} + \sum_{i=1}^{L}\sum_{j=1}^{P}\sum_{j'=1}^{P} y_{ij'}^2$$

$$= 2\sum_{i=1}^{L}[P\sum_{j=1}^{P} y_{ij}^2 - (\sum_{j=1}^{P} y_{ij})^2]. \tag{11}$$

Examined differently, and changing notation for convenience, such that:

$$\sum_{i=1}^{L}\sum_{j=1}^{P-1}\sum_{j'=j+1}^{P} \equiv \sum_{i}\sum_{j}\sum_{j'>j}. \tag{12}$$

Noticing that:

$$\sum_{i}\sum_{j}\sum_{j'}(y_{ij}-y_{ij'})^2 = \sum_{i}\sum_{j}\sum_{j'<j}(y_{ij}-y_{ij'})^2$$

$$+\sum_{i}\sum_{j}\sum_{j'=j}(y_{ij}-y_{ij'})^2 + \sum_{i}\sum_{j}\sum_{j'>j}(y_{ij}-y_{ij'})^2 \tag{13}$$

and since:

$$\sum_{i}\sum_{j}\sum_{j'=j}(y_{ij}-y_{ij'})^2 = 0 \tag{14}$$

then:

$$\sum_{i}\sum_{j}\sum_{j'}(y_{ij}-y_{ij'})^2 = \sum_{i}\sum_{j}\sum_{j'<j}(y_{ij}-y_{ij'})^2 + \sum_{i}\sum_{j}\sum_{j'>j}(y_{ij}-y_{ij'})^2 \tag{15}$$

so, by symmetry:

$$\sum_{i}\sum_{j}\sum_{j'}(y_{ij}-y_{ij'})^2 = 2\sum_{i}\sum_{j}\sum_{j'>j}(y_{ij}-y_{ij'})^2. \tag{16}$$

Combining (11) and (16):

$$2\sum_{i}\sum_{j}\sum_{j'>j}(y_{ij}-y_{ij'})^2 = 2\sum_{i}[P\sum_{j} y_{ij}^2 - (\sum_{j} y_{ij})^2] \tag{17}$$

so that:

$$\sum_i \sum_j \sum_{j'>j} (y_{ij} - y_{ij'})^2 = \sum_i [P \sum_j y_{ij}^2 - (\sum_j y_{ij})^2]. \tag{18}$$

Since, for $y_{ij} \in \{0,1\}$, the left hand side of the theorem:

$$\sum_{i=1}^{i=L} \sum_{j=1}^{j=P-1} \sum_{j'=j+1}^{j'=P} |y_{ij} - y_{ij'}| = \sum_i \sum_j \sum_{j'>j} (y_{ij} - y_{ij'})^2, \tag{19}$$

so that combining (10), (18) and (19):

$$\sum_i [P \sum_j y_{ij}^2 - (\sum_j y_{ij})^2] = P \sum_i \sum_j y_{ij}^2 - \sum_i (\sum_j y_{ij})^2 \tag{20}$$

shows that the pair-wise Hamming distance is equal to the moment of inertia around the centroid times the population size.

4.1 Explanation and Example

The moment of inertia computational method for computing pair-wise Hamming distance works because all coordinates are either 0 or 1. This means that $x^2 = x$ and x times x' is equal to x or x' or both. As a simplified example of how this computational method is used, consider a population of six strings ($P = 6$), each three bits long and having values $y_{gene,individual}$ equal to:

$$y_{11} = 1, y_{21} = 1, y_{31} = 1$$

$$y_{12} = 0, y_{22} = 0, y_{32} = 0$$

$$y_{13} = 1, y_{23} = 1, y_{33} = 0$$

$$y_{14} = 1, y_{24} = 0, y_{34} = 0$$

$$y_{15} = 0, y_{25} = 1, y_{35} = 0$$

$$y_{16} = 1, y_{26} = 0, y_{36} = 1.$$

The coordinates of the population centroid are:
$C_1 = \frac{4}{6} = \frac{2}{3}$, $C_2 = \frac{3}{6} = \frac{1}{2}$, $C_3 = \frac{2}{6} = \frac{1}{3}$.
The population size times the moment of inertia around the centroid

$$P[\sum_{i=1}^{i=L} \sum_{j=1}^{j=P} (y_{ij} - c_i)^2] \tag{21}$$

is computed as:

$$6[(1-\frac{2}{3})^2 + (0-\frac{2}{3})^2 + (1-\frac{2}{3})^2 + (1-\frac{2}{3})^2 + (0-\frac{2}{3})^2 + (1-\frac{2}{3})^2$$

$$+(1-\frac{1}{2})^2 + (0-\frac{1}{2})^2 + (1-\frac{1}{2})^2 + (0-\frac{1}{2})^2 + (1-\frac{1}{2})^2 + (0-\frac{1}{2})^2$$

$$+(1-\frac{1}{3})^2+(0-\frac{1}{3})^2+(0-\frac{1}{3})^2+(0-\frac{1}{3})^2+(0-\frac{1}{3})^2+(1-\frac{1}{3})^2]$$
$$=6(\frac{12}{9}+\frac{6}{4}+\frac{12}{9})=(8+9+8)=25$$

which is the same value as the pair-wise Hamming distance for this population.

The computational efficiency of the moment of inertia method of computing pair-wise Hamming distance makes a considerable difference at population sizes normally encountered in evolutionary computation. For a bit string length of 50 and a population size of 1000, the number of computations necessary for calculation of the pair-wise Hamming distance by the moment of inertia method is two orders of magnitude less than that required by usual computational methods. Even adjusting for the fact that the moment of inertia method involves floating-point calculations, whereas Hamming distance calculations can be made using integer or binary data types, the moment of inertia method for computing pair-wise Hamming distance is considerably more efficient.

5 Relationship to Diversity Measures in Phenotypic Space

For an individual dimension, the moment of inertia measure is closely related to the calculation of statistical variance:

$$\sigma^2=\frac{\sum(X-\mu)^2}{N} \tag{22}$$

differing only in the use of population size in the calculation.

The moment of inertia diversity measure, therefore, when applied in phenotypic space for real numbered parameters is related to the sum of the column variances. It should be noted, however, that when using the moment of inertia population diversity measure for real-numbered parameters, just as when combining traditional column-wise phenotypic diversity measures across columns, attention must be paid to individual parameter scaling. When searching a space, it is important to realize the impact of search-space size on the problem to be solved and understand the resolution (granularity) with which the search for a solution is to be conducted. For example, in a real-numbered convex-space optimization problem, the search space is defined by the ranges of real-numbered parameters. If the range of parameter A is twice as large as that of parameter B, at the same granularity of search, the search space is twice as large along dimension A as along dimension B. In different cases the resolution of interest might be defined as a single percentage of the range, and this percentage might be equally applicable to all parameters. In this case, all parameters should be scaled equally. The moment of inertia calculations can be transformed to equally scale all parameters merely by dividing all parameter values by the parameter range. As long as the parameters are scaled so that they have an equal granularity of interest, the moment of inertia calculations provide an efficient method for measuring population diversity.

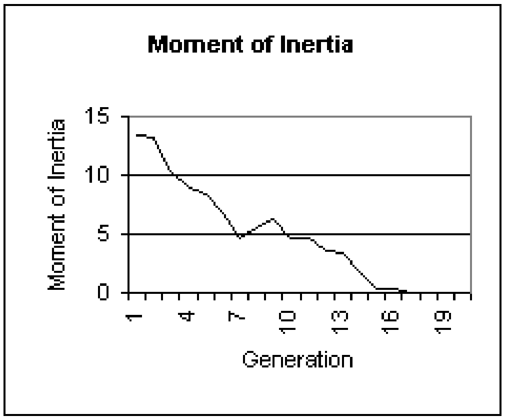

Fig. 8. Moment of Inertia Diversity

It is possible to envision circumstances where it would be desirable to compare the diversity of two different-sized populations on the same problem. In these cases, scaling the diversity by the population size would then be appropriate. When scaled in this manner, the moment of inertia diversity measure for the real parameter problem is equal to the sum of the column-wise variances of the individual parameters.

Figure (8) shows the moment of inertia diversity measure for the example problem used for Figures (1) through (7). Comparing Figure (8) to Figure (6) illustrates that, in addition to being more computationally efficient than the pair-wise column distance measure, the moment of inertia measure more dramatically portrays the population loss of diversity by generation 16 than does the pair-wise distance measure.

6 Conclusions and Future Work

We have introduced a unified method for computing population diversity that is equally useful for measuring diversity for both real-parameter populations and binary populations. Closely related to variance for real-parameter populations, and pair-wise Hamming distance for binary populations, moment of inertia diversity provides a single method of computing population diversity that is computationally more efficient than normal pair-wise diversity measures for medium and large sized EA problems.

The insight into the measurement of population diversity presented here leads to further questions and opens opportunities for other investigation. One area for further investigation relates to whether a suitable Levenshtein-distance [5] version of moment of inertia diversity measurement could be derived, to create more computationally efficient methods of measuring diversity in populations of unequal string lengths. Another area for further research relates to the use of this measure for investigating EA performance. For example, if the population points in N-space are not equally weighted, but are provided "mass" in accordance with their fitness, could the fitness-weighted moment of inertia (the "detected" fitness landscape) and the population moment of inertia (the EA's response to the landscape detection) be used as a measure of EA performance? These and other questions await further investigation.

References

1. DeJong, K.: An Analysis of the Behaviour of a Class of Genetic Adaptive Systems. Ph.D. Thesis, University of Michigan (1975)
2. Mori, N., Imanishi, S., Kita, H., Nishikawa, Y.: Adaptation to Changing Environments by Means of the Memory Based Thermodynamic Genetic Algorithm. In: Proceedings of the Seventh International Conference on Genetic Algorithms, Morgan Kaufmann, (1997) 299-306
3. Horn, J.: The Nature of Niching: Genetic Algorithms and the Evolution of Optimal, Cooperative Populations. Ph.D. Thesis, University of Illinois-Champaign (1997)
4. Morrison, R. De Jong, K.: A Test Problem Generator for Non-stationary Environments. In: Proceedings of Congress on Evolutionary Computation. IEEE (1999) 2047-2053
5. Sankoff, D. Kruskal, J., Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison. CSLI Publications, Stanford, California, (2000)