Distance Measures in Genetic Algorithms*

Yong-Hyuk Kim and Byung-Ro Moon

School of Computer Science & Engineering, Seoul National University
Shillim-dong, Kwanak-gu, Seoul, 151-742 Korea
{yhdfly, moon}@soar.snu.ac.kr

Metric is one of the fundamental tools for understanding space. It gives induced topology to the space and it is the most basic way to provide the space with topology. Different metrics make different topologies. The shape of the space largely depends on its metric. In understanding genetic algorithms, metric is also basic and important. In genetic algorithms, a good distance measure not only helps to analyze their search spaces, but can also improve their search capability. Hamming distance has been popular in most researches for genetic algorithms that deal with discrete spaces. It has also been widely adopted in studies about the analysis of the problem space. In this paper, we propose more reasonable distance measures depending on situations in the process of genetic algorithms and show that they are actually metrics. We propose three distance measures: one for the population-based search, another for the solution space based on K-ary encoding, and the third as an approximate measure of performance improvement of linkage-based genetic algorithms.

Since the genetic algorithm is a population-based search, the distance measure between populations is useful for understanding the behavior of genetic algorithms. We propose an intuitive and reasonable metric.

Definition 1 *Let K be the population size. Let population $\mathfrak{p} = \{\mathfrak{c}_1, \mathfrak{c}_2, \ldots, \mathfrak{c}_K\}$ and population $\mathfrak{p}' = \{\mathfrak{c}'_1, \mathfrak{c}'_2, \ldots, \mathfrak{c}'_K\}$. Given a metric $\mathfrak{d}$ in solution space, we define the distance D_K of the two populations as follows:*
$D_K(\mathfrak{p}, \mathfrak{p}') := min_{\sigma \in \Sigma_K} \left(\sum_{i=1}^{K} \mathfrak{d}(\mathfrak{c}_i, \mathfrak{c}'_{\sigma(i)}) \right)$ *where σ denotes a permutation.*

Theorem 1 *D_K is a metric in the population space.*

Then, it can be computed by the Hungarian method. In the assignment weight matrix $M = (m_{ij})$ between two populations $\mathfrak{p}$ and $\mathfrak{p}'$, each element m_{ij} represents $\mathfrak{d}(\mathfrak{c}_i, \mathfrak{c}'_j)$. The problem of computing D_K is exactly the problem of finding an assignment (permutation) with minimum summation.

We propose a distance measure for disjoint K-grouping problems. In these problems, since each group is not distinguishable, each solution has $K!$ representations. This makes the Hamming distance between two solutions unrealistic and undermines the effectiveness of crossover operators.

Definition 2 *Let the universal set U be $\{1, 2, \ldots, K\}^N$, where N is the problem size. Given two K-ary encodings $\mathfrak{a}, \mathfrak{b} \in U$ and a metric $\mathfrak{d}$ in U, we define the distance measure d_K for K-grouping problem as follows: $d_K(\mathfrak{a}, \mathfrak{b}) :=$*

* This work was supported by Brain Korea 21 Project. The ICT at Seoul National University provided research facilities for this study.

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 400–401, 2004.

$min_{\sigma \in \Sigma_K}(\mathfrak{d}(\mathfrak{a}, \mathfrak{b}_\sigma))$ where σ is a permutation and $\mathfrak{b}_\sigma$ is a permuted encoding of $\mathfrak{b}$ by σ; i.e., i^{th} element e_i of $\mathfrak{b}$ is transformed into $\sigma(e_i)$.

Theorem 2 *d_K is a pseudo-metric in U. Moreover, suppose that Q is the quotient set of U by relation $\sim$ [1] (i.e., $Q = U/\sim$). Then, (Q, d_K) is a metric space; i.e., d_K is a metric in Q.*

When the metric $\mathfrak{d}$ is the Hamming distance $\mathfrak{H}$, the problem of computing d_K is also formulated as the optimal assignment problem. Hence, it can be computed by the Hungarian method. In the assignment weight matrix $M = (m_{ij})$ between two chromosomes X and Y, each element m_{ij} means $\sum_{k=1}^{N} I(X_k = i, Y_k \neq j)$, where N is the length of chromosome and $I(\cdot)$ is the indicator function. The problem of computing d_K is exactly the problem of finding an assignment (permutation) with minimum summation. Since the normalizations of chromosomes pursue the minimization of genotype inconsistency among chromosomes, the proposed metric is ideal in this line of work.

Theorem 3 *If the metric $\mathfrak{d}$ is the Hamming distance $\mathfrak{H}$, then $d_K(X, Y) = min_{\sigma \in \Sigma_K} \left(\sum_{i=1}^{K} \sum_{k=1}^{N} I(X_k = i, Y_k \neq \sigma(i)) \right)$.*

The first level distance measure is commonly the Hamming distance. Other distance measures can also be used as the first level distance (e.g., normalized Hamming distance). We define the second level distance measure. It is defined from the first level distance. Given the problem instance p, consider the graph G_p representing the first order gene interaction; i.e., representing only gene interactions between a pair of genes. Let A_p be the adjacency matrix of G_p.

Definition 3 *Suppose that there exists the inverse of $A_p \oplus I$. We define the second level distance measure D_p as follows: $D_p(\mathfrak{a}, \mathfrak{b}) := \|(A_p \oplus I)^{-1}(\mathfrak{a} \oplus \mathfrak{b})\|$ where $\oplus$ is XOR operator and $\|\cdot\|$ is a norm derived from the first level distance $\mathfrak{d}$ (i.e., $\|\cdot\| = \mathfrak{d}(\cdot, 0)$).*

Theorem 4 *D_p is a metric.*

The second level distance and its extension are efficiently computed in $O(N^3)$ by a variant of Gauss-Jordan elimination method. In genetic algorithms for graph partitioning, gene rearrangement shows dramatic performance improvement on some graphs. The fitness-distance correlation using the proposed second level distance identified the graphs that benefited most by gene rearrangement in genetic algorithms.

Most previous studies needing distances among chromosomes in genetic algorithms used the Hamming distance. The purpose of this paper is to develop more meaningful distance measures for genetic algorithms. We hope that the proposed metrics are useful for improving GA's search capability and understanding GA's working mechanism.

[1] Given an element $\mathfrak{a} \in U$, since $\mathfrak{d}$ is a metric, there are only $K!$ elements such that the distance d_K to $\mathfrak{a}$ is zero. If the distance d_K between two elements is equal to zero, we define them to be *in relation* $\sim$. Then, the relation $\sim$ is an equivalence relation.