Assessment of the Level of Agreement among Metrics for Comparing Non-Dominated Sets

E. O. Oyetunji

Department of Computer Science, Faculty of Mathematical Sciences
University for Development Studies, Ghana

Abstract
The growing interest in the study of multicriteria scheduling problems has led to the development of a number of approximations algorithms for the problem. This has resulted in the need to assess the performance of these approximation algorithms. To this end, many metrics have been proposed. However, there is concern that some of these metrics may produce conflicting results. This then makes comparison of such approximation algorithms difficult. In this paper, a sample of these metrics were tested on a large number of non-dominated (NDS) sets in order to determine whether there is agreement in terms of their assessment of the best non-dominated set for each instance of the problem. Experimental results show that, to a large extent, there is agreement among the metrics when they are combined in assessing the performances of the non-dominated sets rather than when they are used individually.

Keywords: multi-objectives scheduling, agreement, metrics, performance, jobs, machines

INTRODUCTION
The multicriteria scheduling, unlike the single criterion scheduling, involves optimization of two or more scheduling criteria (also called performances measures or objectives). Also, the solution to multicriteria scheduling is normally not a single value but rather a set of values which are often called Pareto set (Rath and Dehuri, 2006). There are many techniques for combining the criteria (Veldhuizen and Lamont, 2000; Hoogeveen, 2005). The technique in which one criterion (the less important criterion) is being minimized subject to the constraint that the second criterion (the more important criterion) is less or equal to a predetermined constant (for a minimization problem) is called hierarchical minimization. In the second technique, the criteria are combined into a scalar function. Each criterion is assigned a weight (which denotes the relative importance of the criterion) by the Decision Maker (DM). The analyst then constructs solutions that minimize the scalar function. This is called the priori technique.

The third technique is called the interactive (or progressive) technique. In this technique, a set of solutions that represents trade-offs on the values of the criteria is sought. Partial preference information (trade-offs) is provided by the DM, while the analyst constructs solutions based on the DM’s partial preference. The solutions obtained are presented to the DM, who can either accept or reject them. The DM’s preferences can be altered if he/she is not satisfied with the solutions presented by the analyst – in which case the analyst begins the search for a new set of solutions based on the DM’s new preferences.

Again the obtained solutions are presented to the DM. This process continues until the DM is satisfied with the presented solution set. The fourth technique involves constructing a set of compromise solutions. In this case, the analyst designs a solution method to seek the set of compromise solutions (non-dominated sets), and presents the obtained solutions to the DM who selects the one that best satisfies his/her preferences. The implications of each solution are presented to the DM. This technique is called posteriori technique. Because of the inability of the DM to specify preference information in many real life scheduling problems, the posteriori technique is of interest in this paper. Many researchers have proposed a number of approximation algorithms to construct the non-dominated sets (Fonseca and Fleming, 1993; Kamiura et al., 2002; Rath and Dehuri, 2006; Ullah et al., 2008; Jadaan et al., 2009). Because of trade-offs, it’s not easy to evaluate the quality of the non-dominated sets (NDS) produced by these approximation algorithms. To this end a number of performance metrics have been proposed for assessing the quality of the NDS (Oyetunji and Oluleye, 2010a). It is therefore necessary to determine the suitability or otherwise of these metrics in assessing the performance of the NDS.

In this paper, the problem of assessing the performances of the non-dominated sets produced by approximation algorithms for the multicriteria scheduling problems using more than one metric at a time is explored. Because performance metrics, when used individually, often give conflicting results, this has necessitated the need to combine them and then
measure their level of agreement in assessing the performances of the non-dominated sets produced by approximation algorithms for the multicriteria scheduling problems.

The paper is organized as follows: section 1 deals with introduction while the review of the literature is presented in section 2. Materials and methods (which covers selection of metrics, determination of reference point and data analysis) is given in section 3. Results and discussions are presented in section 4 while the paper is concluded in section 5.

PREVIOUS WORK

Ever since the realization of the fact that the total cost of a schedule is indeed not a function of one objective but rather a function of two or more objectives (French, 1982), research efforts on the exploration of bicriteria and multi-criteria scheduling problems have continued to grow (Nagar et al., 1995; Stein and Wein 1997; Aslam et al., 1999; Hoogeveen, 2005; Oyetunji and Oluleye, 2008; Oyetunji and Oluleye, 2010a). Many of these research efforts have been directed towards the development of algorithms to obtain the compromised solutions also known as non-dominated sets (Li, 2005; Aittokoski and Miettinen, 2008). Comparing these non-dominated sets is not trivial at al. (Zitzler et al., 2000, Oyetunji and Oluleye, 2010a).

Fonseca and Fleming (1993) studied Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. They showed the ability of the Multi-Objective Genetic Algorithm (MOGA) to uniformly sample regions of the trade-off surface. Deb and Jain (2002) proposed two running metrics for measuring the generation-wise dynamics of multi-objective evolutionary algorithms (MOEAs). Grosan and Dumitrescu (2002) carried out a comparison of some multi-objective evolutionary algorithms using five test functions. The algorithms considered include: Strength Pareto Evolutionary Algorithm (SPEA), Pareto Archived Evolution Strategy (PEAS), Non-dominated Sorting Genetic Algorithm (NSGA II), Adaptive Pareto Algorithm (APA).

Kamiura et al. (2002) explored the Multi-Objective Genetic Algorithm with Distributed Environment Scheme (MOGADES) and proposed a parallel genetic algorithm for the problem. They compared their algorithm with SPEA2 (selected from Zitzler et al., 2001) and NSGA-II (selected from Deb et al., 2000). Their experimental results show that their proposed algorithm (MOGADES) is superior to the other methods used (SPEA2 and NSGA-II). Rath and Dehuri (2006) used a multi-objective evolutionary optimization algorithm called non-dominated sorting genetic algorithm (NSGA) to design a complex heterogeneous embedded system.

Vejandla et al. (2008) compared various MOEAs (NSGAII, SPEA, FastPGA) with a modified version of the Kuhn-Munkres Algorithm (KMA). Experimental results show that the MOEAs were able to produce competitive solutions in reasonable computational time. A new method based on the side information and Non-dominated Sorting Evolution Strategy (NSES)-based K-means clustering algorithm for Accent Classification was proposed by Ullah et al. (2008).

Jadaan et al. (2009) used a method combining the new Non-dominated Ranked Genetic Algorithm (NRGA) with a parameterless penalty (PP) approach to devise the search to find Pareto optimal set of solutions. Their experimental results show that the new algorithm (labeled PP-NRGA) continuously finds better Pareto optimal set of solutions.

MATERIALS AND METHODS

Selection of Metrics

It has been established that metrics that measured both the closeness and diversity aspects of quality of non-dominated sets are desired by researchers (Oyetunji and Oluleye, 2010a). Among thirteen metrics selected for study by Oyetunji and Oluleye (2010a), three metrics are known to cover these two essential quality aspect. These are: Generational Distance (GD) metric which was proposed by Van Veldhuizen (1999), Schott’s Spacing (SS) metric which was proposed by Schott (1995) and Inverted Generational Distance (IGD) metric proposed by Villalobos-Arias et al. (2005). Due to their ease of computation compared with IGD, the SS and GD metrics were selected for study. Also, Oyetunji and Oluleye (2010a) proposed a new metric called Modified Generational Distance (MGD). This metric also covered the two essential quality aspects and does not require the knowledge of the reference set/point unlike the GD, SS and IGD metrics. MGD was also selected for study.

Determination of Reference Point

The SS and GD metrics required the knowledge of the reference set/point before they can be computed successfully. Hence, accurate determination of the reference point is crucial. Therefore, procedure for determining the reference point is hereby proposed as follows:

For a minimization problem;

Step 1: Find the least point (closest point to x-axis) out of the points produced by the non-dominated sets
Step 2: Draw a line (parallel to x-axis) from y-axis to pass through this point
Step 3: Find the least point (closest point to y-axis) out of the points produced by the non-dominated sets
Step 4: Draw a line (parallel to y-axis) from x-axis to pass through this point
Step 5: The intersection point of the two lines is taken as the reference point (for a bicriteria problem). This is the point denoted as $\text{Ref}_{\text{max}}$ in Fig. 1

For a maximization problem;
Step 1: Find the farthest point (longest point to x-axis) out of the points produced by the non-dominated sets
Step 2: Draw a line (parallel to x-axis) from y-axis to pass through this point
Step 3: Find the farthest point (longest point to y-axis) out of the points produced by the non-dominated sets
Step 4: Draw a line (parallel to y-axis) from x-axis to pass through this point
Step 5: The intersection point of the two lines is taken as the reference point (for a bicriteria problem). This is the point denoted as $\text{Ref}_{\text{min}}$ in Fig. 1

**DATA ANALYSIS**

In order to assess the suitability or otherwise of the metrics in assessing the performances of the non-dominated sets, two different non-dominated sets (labeled NDS 1 and NDS 2) were randomly generated. Each set consists of thirty (30) points. The non-dominated sets were replicated 1000 times. The case of bicriteria (two objectives) was considered. The values of the two criteria range between 0 and 50 inclusive.

The three metrics (SS, GD and MGD) were applied to the 1000 instances of the non-dominated sets that were generated. The value of each metric as well as time taken to compute each metric was computed for each instance of the problem. Coding was done in Microsoft Visual Basic 6.0. The data (values of the metrics and execution time) was exported to Statistical Analysis System (SAS version 9.2) for detailed analysis. SAS is a very versatile statistical package, and was employed to enable credible conclusions to be drawn from the results. The hardware used for the experiment was a 1.87 GHz P6000 Intel CPU with 4 GB of RAM.

The frequency (FREQ) procedure in SAS was used to carry out both the frequency distribution and cross tabulation. Also, the general linear model (GLM) procedure was used to compute the mean value of the execution time. The test of means was also carried out using the GLM procedure to determine whether or not the differences observed in the mean value of the execution time are statistically significant.

**RESULTS AND DISCUSSIONS**

Since there are three metrics (MGD, GD and SS) and two non-dominated sets (Set 1 and Set 2), we wanted to know how each of the metrics ranked the sets. To this end, if a metric adjudged set 1 to be better than set 2 or vice versa, it is noted and the number of instances in which each set is preferred by each metric is counted and the results of the frequency distribution of the three metrics are shown in Figs 1-3. The MGD metric adjudged Set 1 as better (Table 1) in 502 (50.2%) instances while it adjudged Set 2 as better in 498 (49.8%). In Table 2, the GD metric adjudged Set 1 as better in 503 (50.3%) instances while it adjudged Set 2 as better in 497 (49.7%) instances. Also, in Table 3, the SS metric adjudged Set 1 as better in 534 (53.4%) instances while it adjudged Set 2 as better in 466 (46.6%) instances. Therefore, isolating the metrics one after the other, they all agreed that the non-dominated set 1 is better than the non-dominated set 2. The sharp difference between the metrics (MGD and GD on one hand and the SS metric on the other hand) could be attributed to the fact that both the MGD and GD measured the distance between the non-dominated set and the reference point/set while the SS metric measures the spacing (distribution) of the points within the non-dominated set.

If we consider the metrics together as a whole (i.e. does the three metrics agree simultaneously or there is at least one metric that disagree?), Table 4 show that the three metrics agree in 696 (69.6%) instances while in 304 (30.4%) instances at least one metric disagree. Going further to see this agreement by the Sets, out of the 696 instances in which the three sets agreed, they adjudged Set 1 as better in 367 (52.7%) instances while Set 2 was adjudged as better by the metrics in 329 (47.3%) instances (Table 5). Where the metrics disagree, the percentage of disagreement with respect to Set 1 is less (44.7%) than that of Set 2 (55.3%) (Table 5). This confirms that the three metrics simultaneously prefer the non-dominated set 1 to the non-dominated set 2. When the metrics were paired in twos (i.e. MGD and GD, MGD and SS, GD and SS), Table 6 show that the highest agreement comes from the GD and SS pair (760 (76.3%) instances), followed by MGD and GD pair (218 (21.9%) instances) while MGD and SS pair lag.
behind (18 (1.8% instances). Again, going further to see this agreement by the Sets, out of the 760 instances in which the GD and SS pair agreed, they adjudged Set 1 as better in 400 (52.6%) instances while Set 2 was adjudged as better in 360 (47.4%) instances (Table 7). Also, out of the 218 instances in which the MGD and GD pair agreed, they adjudged Set 1 as better in 93 (42.7%) instances while Set 2 was adjudged as better in 125 (57.3%) instances (Table 7). In the case of the MGD and SS pair, they adjudged Set 1 as better in 8 (44.4%) instances while Set 2 was adjudged as better in 10 (55.6%) instances (Table 7). In terms of the computational efficiency of the metrics, the MGD is the most efficient, followed by SS while the GD lags behind (Fig 2.). The overall mean execution time taken by the MGD, SS, and GD metrics is 0.056, 0.141 and 0.317 seconds respectively (Table 8). When this was subjected to statistical test, it was found out that the mean of execution time taken by the MGD is significantly different from (more efficient) that of the SS and GD metrics at 5% level (Table 9). Also, the mean of execution time taken by the SS is significantly different from (more efficient) that of the GD metric at 5% level (Table 9). This result is expected as the MGD metric does not require the knowledge of the reference point/set while both SS and GD metrics required the knowledge of the reference point thus confirming the submission of Oyetunji and Oluleye (2010a).

### Table 1. MGD Agreement or Assessment of the metrics

<table>
<thead>
<tr>
<th>Overall_Agree</th>
<th>Frequency</th>
<th>Percent</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGD_Agree</td>
<td>503</td>
<td>50.20</td>
<td>1000</td>
<td>100.00</td>
</tr>
<tr>
<td>Set 1</td>
<td>502</td>
<td>50.20</td>
<td>1000</td>
<td>100.00</td>
</tr>
<tr>
<td>Set 2</td>
<td>498</td>
<td>49.80</td>
<td>996</td>
<td>100.00</td>
</tr>
</tbody>
</table>

### Table 2. GD Agreement or Assessment of the metrics

<table>
<thead>
<tr>
<th>Overall_Agree</th>
<th>Frequency</th>
<th>Percent</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD_Agree</td>
<td>503</td>
<td>50.30</td>
<td>1000</td>
<td>100.00</td>
</tr>
<tr>
<td>Set 1</td>
<td>502</td>
<td>50.20</td>
<td>1000</td>
<td>100.00</td>
</tr>
<tr>
<td>Set 2</td>
<td>498</td>
<td>49.80</td>
<td>996</td>
<td>100.00</td>
</tr>
</tbody>
</table>

### Table 3. SS Agreement or Assessment of the metrics

<table>
<thead>
<tr>
<th>Overall_Agree</th>
<th>Frequency</th>
<th>Percent</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS_Agree</td>
<td>534</td>
<td>53.40</td>
<td>1000</td>
<td>100.00</td>
</tr>
<tr>
<td>Set 1</td>
<td>534</td>
<td>53.40</td>
<td>1000</td>
<td>100.00</td>
</tr>
<tr>
<td>Set 2</td>
<td>466</td>
<td>46.60</td>
<td>996</td>
<td>100.00</td>
</tr>
</tbody>
</table>

### Table 4. Overall Agreement among the metrics

<table>
<thead>
<tr>
<th>Overall_Agree</th>
<th>Frequency</th>
<th>Percent</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGD GD and SS all agree</td>
<td>696</td>
<td>69.60</td>
<td>1000</td>
<td>100.00</td>
</tr>
<tr>
<td>At least one disagree</td>
<td>304</td>
<td>30.40</td>
<td>304</td>
<td>30.40</td>
</tr>
<tr>
<td>Frequency Missing = 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5. Overall Agreement of metrics by Non Dominated Sets

<table>
<thead>
<tr>
<th>Overall_Agree</th>
<th>Best_Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>Row Pct</td>
<td></td>
</tr>
<tr>
<td>Col Pct</td>
<td></td>
</tr>
<tr>
<td>Part_Agree</td>
<td>Frequency</td>
</tr>
</tbody>
</table>

### Table 6. Partial Agreement among the metrics

<table>
<thead>
<tr>
<th>Part_Agree</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only MGD &amp; GD agree</td>
<td>218</td>
<td>21.89</td>
</tr>
<tr>
<td>Only MGD &amp; SS agree</td>
<td>18</td>
<td>1.81</td>
</tr>
<tr>
<td>Only GD &amp; SS agree</td>
<td>760</td>
<td>76.31</td>
</tr>
</tbody>
</table>

### Table 7. Partial Agreement of metrics by Non Dominated Sets

<table>
<thead>
<tr>
<th>Part_Agree</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only MGD &amp; GD agree</td>
<td>218</td>
<td>21.89</td>
</tr>
<tr>
<td>Only MGD &amp; SS agree</td>
<td>18</td>
<td>1.81</td>
</tr>
<tr>
<td>Only GD &amp; SS agree</td>
<td>760</td>
<td>76.31</td>
</tr>
</tbody>
</table>

Frequency Missing = 4
Fig. 2. A Plot of execution time (secs) by metrics by problem instances

Table 8. Overall mean of execution time (secs) by metrics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG_D_Time</td>
<td>1000</td>
<td>0.0562215</td>
<td>0.0346511</td>
<td>0.1133000</td>
<td></td>
</tr>
<tr>
<td>SS_Time</td>
<td>1000</td>
<td>0.1414467</td>
<td>0.0803075</td>
<td>0.2773000</td>
<td></td>
</tr>
<tr>
<td>GD_Time</td>
<td>1000</td>
<td>0.3165402</td>
<td>0.1054864</td>
<td>0.4648000</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Test of means (probability values) of mean of execution time

<table>
<thead>
<tr>
<th>Metrics</th>
<th>MG_D</th>
<th>GD</th>
<th>SS</th>
<th>MG_D</th>
<th>GD</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * indicates significant result at 5% level; Sample size = 1000

CONCLUSIONS

The problem of assessing the performances of the non-dominated sets produced by approximation algorithms for the multicriteria scheduling problems using more than one metric at a time has been explored. A number of metrics were selected from the literature. One major requirement of some of these metrics is the knowledge of the reference point/set, a non-trivial task. In this paper, the procedure for accurate determination of the reference points (for both minimization and maximization problems) have been proposed.

The three selected metrics (SS, GD, and MGD) were tested separately, in pairs and simultaneously. Even though the metrics agreed that the non-dominated set 1 was better than the non-dominated set 2 when they were tested separately, a much more convincing assessment came when the metrics were tested in pairs and simultaneously (all the three put together). Therefore, researchers are encouraged not to use only one metric but a combination of metrics in assessing the performances of the non-dominated sets produced by multicriteria scheduling algorithms. Also, when pairing the metrics, the combination of SS and GD metrics have the tendency to agree better than others in adjudging the best non-dominated set.

REFERENCES


