ANSWERS TO REVIEWERS

Valentino Santucci, Marco Baioletti, Alfredo Milani

REVIEWER 1

Compared to the RCRA version, the article has been improved to give more details and precisions about the differential mutations. A second extension is the experiments on the version of the problem with makespan minimization where very good results are reported. In the end, the article describes well the approach and the reported results are even more solid than in the workshop version.

But I have the same main concern as for the workshop version about the lack of evidence in the experiments that DE really brings an added value in the overall approach.

> Criticism: In fact, my main comment and question about the paper is that, although differential
> mutation seems to be an interesting and original tool in this type of problem, the experiments do not
> really provide an evidence for that. The experiments seem to clearly indicate that the overall approach
> performs very well on the problem, but the differential evolution is only one of the ingredients of the
> approach beside the crossover, the way selection is performed, the use of a constructive heuristic in
> the initial population, the local search, the restart, etc. It would for instance be very informative to see
> how the approach behaves when differential mutation (in line 4 of algorithm 2) is replaced by, let's say,
> the random selection of an individual in the population or a purely random individual, or the result of a
> path relinking between two randomly selected individuals, ... Otherwise, in the end, even if the idea of
> differential mutation looks attractive, there is no real proof that it is the main ingredient to the success
> of the approach.

>> Answer: It worths to note that: (1) AGA (see paper references) employs our same crossover and their
>> performances seems to be inferior, (2) population restart should be regarded as a sort of endemic
>> component of differential evolution (DE); indeed, DE population is not able to evolve when all the
>> individuals are the same, also in the original and continuous DE, (3) hybridization of population-based
>> algorithms with local searches is a widely exploited technique in the field (see for example AGA or
>> HGM-EDA), (4) the same is true for the construction of initial solutions by means of heuristic methods
>> (see for example AGA or IG). Summarizing, local searches, restarts, and guided initialization are widely
>> adopted in any other state-of-the-art algorithm.

(1) Yes, but is the difference between AGA and the approach described in the paper only due to the DE crossover ? I don't think. And even if that was the case, due to the many implementation details, the only way to measure the impact of the DE component of the approach is I think to try to replace it with a more naive component and perform a comparison in the same implementation framework
(2-4) I know that restarts, hybridization with LS, construction of initial solutions with heuristics, etc are widely used in this type of approaches. I do not question the fact that the proposed approach
needs all these ingredients to be competitive, this is to be expected.
My point is just that the article focuses on Differential Evolution algorithms (which is a great idea because as mentioned these methods are not very used in discrete optimization) but the experimental study does not really show that *this* ingredient helps.

>> Anyway, we are working on an experiment that allows to objectively made more clear the contribution
>> of our differential mutation.

Indeed, that would definitively help improving the article. I think it should not be very difficult to rerun the experiments using the same approach but just replacing the crossover operator by a simpler one that do not use DE and compare the results to demonstrate the added value of DE.

Answer 1: We added an experimental section in which we compare our DE with other 3 versions in which the differential mutation we have proposed is replaced by:
1) a totally random mutation
2) a random individual from the population
3) an individual created by means of path-relinking
Moreover, for the sake of clarity, it worths to note that Differential Evolution is the name of the main evolutionary scheme that includes (and organizes) the three genetic operators of differential mutation, crossover, and selection. Therefore, the differential mutation is only one operator of Differential Evolution and, although it builds a mutant by using other population individuals, it is generally considered a mutation operator and not a crossover operator.

On a similar line, as the complexity of the randomized bubble sort is in O(n^2), is it not expensive on large instances? As far as I see, the comparisons are based on a fixed number of iterations or objective function evaluation. On PFSP, objective function evaluation is in O(n) so the number of evaluation, at least in case of DE, does not seem to give a good evaluation of the running time of the approaches. I know it is difficult to compare with running times due to different machine performance, implementation effort, programming language, etc. At least, if you study the impact of using DE or a more naive approach for the crossover in your implementation framework, as you are working in the same framework and same machines, a comparison using running times would make sense.

Answer 2: In the general case, the complexity of the fitness computation is O(nm). We added a new section which compares the execution times of DEP with the other competitor algorithms. It worths to note that the local search based algorithms can exploit a technique that allows to speed-up the fitness computations of multiple neighbors of an incumbent solution. Unfortunately, this technique is not applicable to population based schemes like DEP. However, as described on the new Section 6.5, DEP is faster than the population-based competitors, i.e. GM-EDA and HGM-EDA.

I think the reported results of the overall approach are extremely good and I don't doubt that Differential Evolution is a key element of the approach. So my recommendation is to resubmit a revision of the article that includes the ongoing experiments that you mention in the accompanying letter so that the interest of using DE in the approach is clearly demonstrated.

Answer 3: Done.

REVIEWER 2

The article proposes a new crossover operator for evolutionary algorithms, that can be used in problems where the solution is represented by a permutation. Tests on the contribution of the new operator are conducted only on the flow shop scheduling problem under two different optimization criteria. The results presented are promising.

Although interesting and suited for a conference venue, I do not deem the article ready for publication in a journal. The following are the main reasons for this assessment:

- In the conclusions the authors claim that the operator is suitable for other permutation problems as well, TSP, QAP, LOP, and that they are going to try in future research. Although the application to each of these problem may lead to a conference paper by itself, I think that a journal paper should present the collection of these results and a more general assessment.

Answer 4: We think that trying our DE algorithm to each of these problems should be seen as a paper by itself. See for instance the large literature where optimization algorithms for permutations space are applied to only one problem (e.g., references [5,6,10,11] of the paper).

- The problem in focus is an easy to formulate problem that has been well studied in the literature. I am strongly convinced that anyone developing a novel algorithm for this problem must publish the source code. Although it is has not been the practice in the previous years, the situation must change or we will be here in ten years from now still dealing with papers that claim to have new best result on well known benchmarks without being able to trust the results (see the next point). Note that working conditions make the sharing of source code much easier than it was say 10 years ago.

Answer 5: We have just published the code on github at the following url: https://github.com/goldengod/dep

- The algorithm designed uses several well known algorithmic components and add a couple of new features. The claim is that the algorithm is able to find the best known results on the makespan criterion, which is the most studied version. But I do not think the authors bring enough (empirical) evidence that these results are achieved thanks to the new components and not thanks to the previously designed ones.

The focus of the paper is on the new operator. The goal is to show that it leads to improved performance. However, in the analysis the contribution of this component is not separated from the rest, and rather mixed with other elements, which make difficult to see whether is actually responsible for the improvement.

Answer 6: to the best of our knowledge, the proposed DE algorithm for permutations is original: the mutation and selection are new, while only the crossover has been chosen among the best crossover operators for PFSP. Moreover, we have performed some additional experiments (in Section 6.1 and 6.4) to show the important contributions of the newly proposed mutation and selection operators.

- The design of the experiments is not clear enough. As far as I know there has been more focus on the makespan objective for this problem. When presenting the results in Section 4.2, the algorithm DEP is compared with IG, GM-EDA, HGM-EDA. It is not clear why, it is not compared with VNS, and why the results and settings of reference [4] are not taken into account. Further, in section 4.1 VNS assesses solutions by delta evaluation, which is its strength, hence it is not fair to terminate the algorithm on the basis of a limit on the number of evaluations, since for VNS they are less costly. A similar reasoning probably holds for IG in section 4.2 although there, it is not even stated what has been the termination criterion used. More in general, I do not think we can remove the computation time from the analysis, as done in the paper, without runninng the risk of being unfair in the conclusions.

Answer 7: The design of experiments should now be clear. VNS has been used only for PFSP-TFT, because it is not among the state-of-the-art for PFSP-Makespan. Reference [4] is outdated and IG was proposed later in [6]. Finally, we have added Section 6.5 in which we compare the execution times of the algorithms used in the paper. Anyway, the criterion based on the number of the fitness evaluations is widely adopted in the community of evolutionary computation and we think that using only the computation time as the stopping criterion would be unfair with respect to our algorithm and other evolutionary methods. See also Answer 2.

- The authors claim that the proposed operator extends the "contour matching" property of DE from continuous to discrete optimization. However I could not find a definition of this property and an explanation of this claim in the paper and I cannot find one myself. The whole derivation from differential evolution seems to me forced and more confusing than needed. Or anyways the motivation for involving Differential Evolution and for doing the effort of adapting to the discrete case a global optimization algorithm, is not explained. I am not aware of any idea behind algorithms for continous optimization that has been succesfully applied to discrete optimization. If the authors want to do this, then I expect a motivating explanation.

Answer 8: Contour matching is defined in the popular book “Differential Evolution: A Practical Approach to Global Optimization” (at page 44) but it is an informal definition and therefore we have decided to remove the reference to the “contour matching” property to avoid possible misunderstandings.

- There is a contradicting claim in the article, on page 3 it is said that "spaces of smaller diameters are like those with generators T and I would have a reduced number of possible values and we expect that it would produce worse results". Later on page 6 it is said that: "the weakness of DEP is probably due to a very slow convergence of the DEP population caused by the extremely large diameter of the search space where the differential mutation navigates".

Answer 9: we have removed that explanation. Actually, the cause is likely to be the very slow evolution.

- There has been a large focus on tuning methods in the last decade. I do not think that in a journal paper it is enough to say that the value of parameters was chosen out of preliminary experiment. A sound methodology has to be used and details of the application given. In particular it must be clear what range of value were considered.

Answer 10: we added Section 6.1 describing in a full detail the tuning procedure we used.

- There are too many external references that make it impossible to read and understand the article independently from previous knowledge on the problem.

Answer 11: we tried to make the paper more self-contained by introducing new descriptions and examples. Anyway, it is very hard to make it completely self-contained without losing the focus on the main contribution of the paper.

Other comments:

- How are the optimal results known? Are they known for all the benchmark instances?

Answer 12: the additional material of reference [5] (see the url http://www.sc.ehu.es/ccwbayes/members/jceberio/GMallowsEDA/GMEDA.html) comprises:

- ***best and average TFT value obtained by AGA, VNS4, GM-EDA and HGM-EDA,***
- ***best and average makespan value obtained by GM-EDA and HGM-EDA.***

The best values considered are taken as the minimum among these results and the results we have obtained by our executions of DEP and IG. Moreover, in order to make a consistent experimental investigation, the same budget of fitness evaluations and the same number of executions of [5] are used.

- On page 3, the authors use permutation group terminology without defining it or giving reference. What is a generator? What is the meaning of truncating a permutation $F\cdot \pi$? In the definition $F\cdot \pi=g_1 \circ g_2 \ldots g_k$, I lack to see where is the influence of π on the right hand side. Thus, I have a problem with the definiion (5). Further, in the enumeration list that follows below, the terms "difference between permutations", "scaled difference" and "truncated shortest path" they seem all to be left undefined to me. As well as later the term "diameter". Describing the

Kendal-τ distance I think the authors miss to say that the number of inversions are needed to bring $\pi^-1\circ \pi_1$ to the canonical permutation. At the end of page 3, what is e in $d_K(\pi,e)$? I think I understood what the authors are doing with this operator and I know the reference [22]. I think that it is a nice idea! But I also mean that all this should be explained better. A numerical example for all these operations in this part of the paper would be very helpful!

Answer 13: We added Section 3, which provides a short introduction to group theory and permutation groups. Moreover we added an example of mutation. These should clarify whhich are the background concepts and which is our contribution.

- In section 3.2 what is the initial value of F_i?

Answer 14: the initial value of F is 0.5. However, this value is automatically and continuosly adapted by the jDE scheme (see Section 5.2), thus its initial value has a negligible contribution.

- At the end of section 3.4, it is not easy to see the logic behind point 2). If there is a desire to avoid restart then simply avoid introducing it in the algorithm tout court. Or change the criterion for activating restart. Reason 1) is enough. In section 3.5, I do not understand what should make an individual the "best" when they have all the same fitness.

Answer 15: We have modified this part, but it is important to understand that (as explained in Section 5.5) a restart mechanism is strictly necessary.

- What is the computational cost of a single move in the local search?

Answer 16: Each local search loop, in the worst situation, examines all the neighborhoud of the incumbent solution. As can be easily shown, there are O(n^2) neighbors both for the insertions and interchanges neighborhouds. Moreover, each neighbor requires O(nm) time to be evaluated. However, multiple neighbors evaluation can be sped-up using the procedure described in [7]. This is also discussed in Section 6.5 and it is worthwhile to note that this speed-up technique has been implemented in all the local search codes used in the experimental session for the execution time.

- In Section 4, towards the end of page 5, what are n and m in $n\times m$ problem configurations?

Answer 17: As described in the introduction, n is the number of jobs and m is the number of

machines of a PFSP instance.

- At the end of page 5, second column, I assume that 0.05 was the significance level used and not the confidence level. Else it would be a rather wired choice.

Answer 18: Done.

- In the analysis of experiments were there taken into account 20 results per instance also for the other algorithms compared? Was the "best" value in the ARPD calculation the overall best also including the results of this article or only the previous best known? If it was the former then I would like to have a clarification about how ARPD is calculated for the other algorithms, if it is the latter then I would expect some negative results in the ARPD reported in the table.

Answer 19: The ARPDs are computed by using the best values found as described in Answer 12 and using equation (9). In the cases where we have not run the algorithms (i.e., AGA, VNS4, GM-EDA, HGM-EDA for PFSP-TFT, and GM-EDA, HGM-EDA for PFSP-Makespan), we have their average results as described in Answer 12. Now, a bit of algebra allows to express the ARPD formula of equation (9) as a function of best and average values. Therefore, our experimental comparison is consistent and sound.

- When commenting results on the basis of the best results, please be aware of the pitfall described in this paper: M. Birattari and M. Dorigo (2007). How to assess and report the performance of a stochastic algorithm on a benchmark problem: Mean or best result on a number of runs? Optimization Letters, 1(3):309-311.

Answer 20: We described both best and mean results. Indeed, the ARPD values are a sort of normalized average results of the algorithms considered.

- It is not clear why in section 4.1 the authors distinguish number of jobs and in section 4.2 distinguish number of machines. These are factors in the analysis and the evidence of their relevance has to be shown by means of the experimental results. Not ad libitum.

Answer 21: The different discussions about the experimental sessions for PFSP-TFT and PFSP-Makespan are motivated by the evidence, as shown in Tables 4 and 5, that on PFSP-TFT the performances of the competitor algorithms seems to vary more with the increasing of n with respect to m. Instead, the converse looks to be evident on PFSP-Makespan.

- The claim that a randomly biased selection operator allows to improve the population diversity should be substantiated by results. For example one could show that the number of restarts change or results overall improve. This is not shown in the paper.

Answer 22: This claim has been modified (see Section 5.3). The intuitive explanation does not need a deep experimental investigation. Indeed, it is evident that a way to stop population stagnation (that is, no change in the population) is to accept worsening individuals, thus allowing the evolution to resume.

- In a few places, I suggest the adoption of more common terms. Thus, flow time is known as completion time (see text books on scheduling such as Pinedo, M. Scheduling: Theory, Algorithms, and Systems Springer New York, 2008 and Brucker, P. Scheduling Algorithms. Springer 2007). Neutrality is known as large plateux (see indeed reference 29). Names are arbitrary of course, however, in scientific fields, it helps communication to refer to the same concepts with the same names.

Answer 23: The more used term is total flowtime as in the paper (see for instance [5,10,11,12]). The neutrality term (though correct in the fitness landscape analysis community) has been replaced by "more and larger plateaus".

Together with the paper, you received its workshop reviews as well as a letter
from the authors explaining how they have dealt with these
reviews in the current version.

If you already reviewed this paper at the workshop: did authors properly take into
account your previous comments?
(Please, delete all categories except one)

- I did not review this paper at the workshop.

Did authors properly take into account the comments of the reviewers (other than you)?
(Please, delete all categories except one)

- No.

If you are not fully satisfied about how authors addressed reviewers' comments, please explain why:

The main comment of Reviewer 1, which is also one of mines above, has not been addressed. Authors claim they were working at an experiment but I do not see it in the paper.

Several other comments about expanding the explanation and adding charts that clearly visualise the results have been completely ignored in the paper as well as in the replies.

====== END ==========================

REVIEWER 3

Together with the paper, you received its workshop reviews as well as
a letter
from the authors explaining how they have dealt with these
reviews in the current version.

If you already reviewed this paper at the workshop: did authors
properly take into
account your previous comments?
(Please, delete all categories except one)

- I did not review this paper at the workshop.

Did authors properly take into account the comments of the reviewers
(other than you)?
(Please, delete all categories except one)

- There are still some important points that have not been addressed, neither in the paper nor in the accompanying letter.

If you are not fully satisfied about how authors addressed reviewers' comments, please explain why:

this paper introduces a new discrete differential algorithm for solving permutation flowshop scheduling problems. A new mutation scheme is proposed which is based on a new randomized bubble sort algorithm. This paper is essentially the same paper that has been accepted to the RCRA workshop.

Answer 24: This workshop paper has been extended here by introducing: an experimental evaluation on the makespan objective, an experimental calibration of the parameters, an experimental discussion about the impact of the newly proposed differential mutation, a parametrization of the selection operator, more descriptions and more sections.

Not being an expert on the topic, my attention got focused on the experimental evaluation. The principal claim of the authors is that for 45 instances (out of 120) is found a new best value. Although this is significant in terms of results, the authors do not conduct a study to better understand why this happens. A more thorough analysis of the results obtained is needed.

Answer 25: This study is provided both in Section 6.1 and 6.4.

In the workshop paper is mentioned that you are planning to extend the use of the DEP algorithm to other permutation-based problems, for which a few examples are given. This is not done in the journal paper but actually should be. Indeed, the results are exactly the same in the two publications.

Answer 25: The workshop paper has been extended by adding an experimental investigation for the PFSP with makespan objective. As described in Section 6, PFSP-TFT and PFSP-Makespan, though looking similar problems, have very different characteristics. Regarding the application of DEP to other permutation-based problems, see Answer 4.

Minor comment:

It is unclear whether the authors have run the other solvers on their machines. Otherwise the comparison is not fair.

Answer 26: The comparison is fair and sound. See the Answers 12 and 19.

====== END ==========================

REVIEWER 4

The paper proposes an interesting (and rather complex) method for the classical flow-shop problem, with two different objective functions.
In addition, the experimental results are rather good and compared to the state-of-the-art in a statistically-principled way. Overall, it is a good paper.

I have only a couple of comments:

1. results are compared based on the number of evaluations, as suggested in a previous work. Nevertheless, I think that running times should also be reported, as the number of evaluations might not has a unique meaning throughout different techniques.

Answer 27: Execution times are provided in the newly introduced Section 6.5. Regarding the termination criterion, see Answer 7 and 2.

2. I believe that the same statistical rigor used for the comparison with the literature should be used also for the parameter tuning phase, which instead is done in a rather naive way (if I understood correctly).

Answer 28: An statistically sound experimental parameter tuning is now introduced in Section 6.1.