Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-12 (2016) 431–436

Data Mining Approaches for the Methods to Minimize Total Tardiness in Parallel Machine Scheduling Problem

Ozlem Senvar*, Farouk Yalaoui, Frédéric Dugardin, , Andres Felipe Bernate Lara

* European Commission post doctoral researcher grantee Ref.2013/327-350

Laboratoire d'Optimisation des Systèmes Industriels (LOSI),
Institut Charles Delaunay (ICD), UMR CNRS 6281,
Université de Technologie de Troyes,
12, rue Marie Curie, CS42060, 10004 Troyes cedex - France

Abstract: This study examines large sample size of instances and tries to extract useful knowledge about the domain of parallel machine scheduling problem (PMSP) and solution space explored. The aim of this study is to provide statistical interpretations and classify the differences between the instances. The interrelationship between specified predetermined inputs and the output is examined through artificial neural networks (ANNs) along with regression analysis since they can easily explore which inputs are related to the output and develop regression model. The results of both analyses reveal significancies of the relationships and predicted importance of the predetermined inputs on the output. Furthermore, we examined the behaviours or patterns of the instances, after realizing the easiness and hardiness of the instances accentuating the differences. In order to link the predetermined inputs of instances with the performances of the set of tested methods, the differences between instances are evaluated in terms of variability. Then, we grouped instances into three clusters, specifying as exact, equal and difficult zones, for information retrieval about their complexities via hierarchical clustering method.

Keywords: Parallel machine scheduling problem (PMSP), Total Tardiness, Artificial Neural Networks (ANNs), Regression Analysis, Agglomerative Hierarchical Clustering.

1. INTRODUCTION

This study handles the PMSP of a set of N jobs on M identical parallel machines in order to minimize the total tardiness over all the considered load without any preemption or splitting. Each job has its own processing time, release date and due date. All the machines are considered identical (with same speed) and available during all the scheduling period. Our problem is literally denoted as $P_m/r_i/\sum T_i$ corresponding to minimization of the sum of tardiness with a set of N jobs having different release dates (r_i) and M identical parallel machines. Notably, the considered problem is NP-hard since it is harder than the single machine scheduling problem, the relaxation without release date, was proved to be NP-hard by Koulamas (1994). Yalaoui (2012) proposed an exact resolution, an Ant Colony Algorithm (ACO), a Tabu Search (TS) method, a set of heuristics based on priority rules and an adapted Biskup Hermann Gupta (ABHG) method to solve the problem.

In this study, we consider the problem including the sequence decision as in the single machine problem and considering the job assignment into a set of machines. We consider the cases with $M < N$ due to the fact that optimal solution is evident when $M \geq N$. In order to solve the defined scheduling problem, we generated considerably large sample size of instances. We obtained a set of 4500 instances from generator model using uniform distributions in range for release dates (r_i), processing times (p_i) and due dates (d_i) to execute a set of solutions for our proposed 171 methods to minimize tardiness in PMSP. That is, using large size of 4500 instances, a total set of 171 heuristic and metaheuristic methods are proposed and tested for a permutation of initial solutions and iterative search procedures with short computational times. For each instance; predetermined inputs are specified as number of jobs (N), number of machines (M), parameters α and β which are used to generate release dates and due dates, respectively in the instance generator. On the other hand, for each instance; output is specified as mean total tardiness (MTD), which is evaluated by the average of total tardiness of each 171 proposed methods for each instance. We preferred to utilize output as MTD because the mean values are used as measures of location along with accuracy in descriptive statistics. Notably, MTD should not be confused with the mean tardiness, which is generally utilized as due date related indices in scheduling theory.

The aim of this study is to provide statistical interpretations and examine the differences between instances because we realized that easiness and hardiness of the instances accentuating the differences. We evaluated the differences between instances in terms of variability. Notably, from statistical point of view, variability is defined as the quality of non-uniformity of a class of entities or instances. Consequently, this variability causes delay or departure of a production process from its regular behaviour.

To the best of our knowledge, in the literature such kind of study has not been accomplished yet. On the other hand, it

Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.07.645

has to be emphasized that it is practically difficult or impossible to analytically relate such kinds of relationships between the job characteristics and the performance measure of the scheduling system. In this study, the interrelationship between specified predetermined inputs and the output is tried to be explored through artificial neural networks (ANNs) along with regression analysis. By exposing instances for the relationship to learning abilities of the both techniques, complex relationships are captured between the input and output variables. Both ANNs and regression analyses are sometimes utilized as alternative techniques to each other due to the fact that both techniques are particularly useful for dealing with multivariate relationships (Senvar et al., 2013). For our considered problem in which prior knowledge required to facilitate manual tagging about instances is unavailable or insufficient in scheduling domain. For examination of each instance, clustering is more suitable option after supervised learning approaches like ANNs along with regression. ANNs along with regression analyses and clustering are data analysis tools in data mining, machine learning, and neurocomputation.

The rest of the paper is organized as follows: Section 2 gives literature review in brief. Section 3 briefly presents methodology. Section 4 provides application study along with results. Section 5 gives conclusions and recommendations.

2. LITERATURE REVIEW

Total tardiness minimization has not been studied as much as other performance criteria like makespan for PMSPs (Albers, 2013). Yalaoui and Chu (2002) addressed the identical PMSP to minimize total tardiness and developed dominance properties, lower and upper bounding schemes incorporated into a branch and bound algorithm. Dugardin et al. (2007) used an Ant Colony Optimization (ACO) algorithm to minimize the total tardiness for a hybrid flow-shop and PMSP. Bernate-Lara et al. (2012a) presented a tabu search method for dealing PMSP with total tardiness minimization criterion when jobs have different release dates and they had performance comparisons of the solution method with the ABHG method and the local search algorithm.

Golmohammadi (2013) highlighted that previous neural networks models in scheduling focus mainly on job sequencing and simple operations flow, and may not consider the complexities of real world operations. Akyol and Bayhan (2007) also emphasized that after learning the unknown correlation between the input and output data, ANNs can generalize to predict or classify for the cases they were not exposed to.

Hierarchical clustering has great importance in data analytics due to the exponential growth of real-world data, which are unlabelled and there is little prior domain knowledge available. One challenge in handling these huge data collections is the computational cost (Bouguettaya et al., 2015).

Principally, traditional linear regression model can acquire knowledge through the least squares method and store that knowledge in the regression coefficients. However, linear regression has a rigid model structure and set of assumptions that are imposed before learning from the data (Turanoglu et al., 2012). In contrast to traditional linear regression model, ANNs impose minimal demands on model structure and assumptions. A neural network can approximate a wide range of statistical models without hypothesizing in advance certain relationships between the dependent and independent variables. Instead, the form of the relationships is determined during the learning process. If a linear relationship between the dependent and independent variables is appropriate, the results of the neural network should closely approximate those of the linear regression model. If a nonlinear relationship is more appropriate, the neural network will automatically approximate the "correct" model structure (Senvar et al., 2016).

3. METHODOLOGY

The considered problem is theoretically defined and expressed mathematically in (Bernate-Lara, 2014). To solve the defined scheduling problem, a set of heuristic and metaheuristic methods are developed and tested. Totally, 171 different methods are developed for permutation of initial solutions and iterative search procedures. Parameters are established as 12, 40 and 100 jobs, scheduled in 2, 3 and 5 machines. In other words the instances are generated for number of jobs, denoted as $N \in \{12,40,100\}$, and for number of machines denoted as $M \in \{2,3,5\}$.Computational tests for our 171 proposed methods in a set of 4500 instances are based on incoming values for N, M, and framework of processing times (p_i), release dates (r_i), and due dates (d_i), which are generated using uniform distributions in range of instances generator model developed by Yalaoui (2012).This instance generator uses the structure presented as follows: In the instance generator, for each instance; processing times (p_i) are random integer numbers generated from uniform distribution as $p_i \sim U[1,100]$. Let TPr be the sum of processing times of N jobs. Release dates (r_i) are random integer numbers which are generated from uniform distribution as $r_i \sim U[0,\alpha\times \mathrm{TPr}/M]$. Due dates ($d_i$) are random integer numbers which are generated from uniform distribution as $d_i \sim U[r_i+p_i, (r_i+p_i)+\beta\times \mathrm{TPr}/M]$.

Yalaoui (2012) defined parameters α and β for the instance generator as follows: $\alpha \in \{0.2,0.7,1.2,1.7,2.2\}$ and $\beta \in \{0.02,0.27,0.52,0.77,1.02\}$ and identified the impact of α and β values on the complexity of solving any instance. Combination of the parameters defines difficulty of instances.

Table 1 gives the intervals for the parameters α and β where integer numbers are used to identify each α/β couple. Computational tests can be considered in 25 groups according to the number of possible combinations of parameters α and β. It has to be taken into account that, the groups are characterized by the time passed to solve an instance or the number of outperforming results. To sum up, there are groups with short and large time consuming.

For each instance, the standard deviation is computed as follows: