Expert Systems with Applications 42 (2015) 4029–4041

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Bi-criteria appointment scheduling of patients with heterogeneous service sequences

Alireza Saremi [a], Payman Jula [b,*], Tarek ElMekkawy [c], Gary G. Wang [d]

[a] Fraser Health Authority, Surrey, BC, Canada
[b] Beedie School of Business, Simon Fraser University, Vancouver, BC V5A 1S6, Canada
[c] Department of Mechanical and Industrial Engineering, Qatar University, Doha, Qatar
[d] Mechatronic Systems Engineering, Simon Fraser University, Vancouver, BC V3T 0A3, Canada

ARTICLE INFO

Article history:
Available online 13 January 2015

Keywords:
Outpatient and surgery scheduling
Simulation-based optimization
Mathematical programming
Multiobjective Tabu search
Multiagent optimization

ABSTRACT

This article addresses the challenges of scheduling patients with stochastic service times and heterogeneous service sequences in multi-stage facilities, while considering the availability and compatibility of resources with presence of a variety of patient types. The proposed method departs from existing literature by optimizing the scheduling of patients by integrating mathematical programming, simulation, and multiobjective tabu search methods to achieve our bi-objectives of minimizing the waiting time of patients and the completion time of the facility. Through intensive testing, the performance of the proposed approach is analyzed in terms of the solution quality and computation time, and is compared with the performance of the well-known method, Non-Dominated Sorting Genetic Algorithm (NSGA-II). The proposed method is then applied to actual data of a case study operating department in a major Canadian hospital and promising results have been observed. Based on this study, insights are provided for practitioners.

1. Introduction

Advances in healthcare in recent years accompanied several efforts to improve the efficiency of operations in healthcare in order to reduce the expenses in this sector. Accordingly, many studies have been carried out to improve the efficiency of scheduling in order to address the increase in demand for outpatient and inpatient services. In this article, we consider multi-stage facilities that serve patients of different types with non-identical stochastic service time at each stage. We assume heterogeneous service sequences; i.e., each patient type follows a specific order to visit stages of the facility that may vary from type to type. For instance, a surgical patient may go through stages such as reception, pre-operation, operation, and post-operation, while a checkup patient may undergo a different sequence of stages. In this manuscript, scheduling refers to the determination of the arrival time for each patient to the facility in order to minimize bi-objectives of the waiting time of patients and the completion time of the facility.

Relevant literature indicates that analytical methods and simulation studies have been used to solve the problem of appointment scheduling and planning in healthcare, including inpatient and surgical facilities. Typically, optimization methods use analytical approaches to achieve optimal (or near optimal) solutions. These approaches generally have difficulty addressing large and complex systems. On the other hand, simulation methods can address many complexities in large systems. However, simulation methods are time-consuming and often do not deliver a competitive optimization strategy (Cayirli & Veral, 2003). A gap still exists in the literature for efficient and effective methods to address the challenges in scheduling of such services. In this article, efficiency of a method refers to the amount of computation time required by the method to produce results, while effectiveness addresses the quality of solutions generated by the method. One of the contributions of this article is that we target this gap by integrating analytical methods and simulation. Another contribution is that in contrast with commonly used single objective optimization methods, our approach provides a Pareto front for bi-objectives of patients' average waiting time and the facility completion time. The average waiting time refers to the time in which patients have to wait to receive various services in the facility, while facility completion time refers to the time that the last patient leaves the facility. The Pareto front (also known as Pareto set, or Pareto frontier) is the set of choices that are not strictly dominated by another point in the objective space. Finally, we depart from existing literature by considering the

* Corresponding author.
E-mail addresses: Alireza.Saremi@fraserhealth.ca (A. Saremi), pjula@sfu.ca (P. Jula), tmekkawy@qu.edu.qa (T. ElMekkawy), gary_wang@sfu.ca (G.G. Wang).

http://dx.doi.org/10.1016/j.eswa.2015.01.013
0957-4174/

health centers that serve patients with heterogeneous service sequences.

This work proposes an optimization method termed multi-agent tabu search (MATS), which simultaneously addresses bi-objectives of minimizing patients' waiting time and the clinic's completion time. MATS uses mathematical programming (MP), tabu search, and simulation model. The MP model provides MATS with promising initial solutions. Tabu search then improves the initial solution by searching for optimal schedules by running a number of agents in parallel. The agents seek the non-dominated solutions of the problem and share information with each other to improve the search performance of the algorithm. In order to capture the complexity of multi-stage facilities, MATS is performed on a discrete event simulation. Therefore, MATS method benefits from the flexibility of simulation, and the power of mathematical programming optimization to find Pareto fronts.

In order to evaluate the performance of the proposed method, we developed several test problems with a range of important factors such as, the number of patients and patient types, and the coefficient of variation of service times. We compared the performance of MATS with Non-Dominated Sorting Genetic Algorithm (NSGA-II) in terms of quality of solutions and computation time. NSGA-II has been selected for evaluating performances since it is one of the most powerful methods in the field of multiobjective optimization, and has recently been used in healthcare appointment scheduling with promising results (Gul, Denton, Fowler, & Huschka, 2011). To measure the quality of solutions, we use the hyper-volume and spacing performance indicators, which will be explained later in this article.

This paper has been organized as follows: Section 2 presents a literature review for the appointment scheduling problem in surgical and outpatient settings. Section 3 discusses the problem definition. Section 4 describes the architecture of the proposed approach and describes different components of the algorithm. Section 5 gives the design of experiments and analysis of the results. Section 6 provides a case study of an OR department in a major Canadian hospital. Section 7 provides insights for practitioners. Finally, Section 8 provides the conclusions and directions for future work.

2. Literature review

This section presents a brief review of the relevant literature. Here, we consider either articles that study the appointment scheduling problem as a multi-stage facility, or those which use discrete event simulation or mathematical programming in their study. We further divide appointment scheduling into outpatient appointment scheduling and surgery scheduling. For a more comprehensive review of literature, readers are encouraged to refer to Cayirli and Veral (2003), and Gupta and Denton (2008) for general outpatient appointment scheduling, as well as Blake and Carter (1997) and Cardoen, Demeulemeester, and Belien (2010) for surgery scheduling.

We divided the relevant literature into three categories: optimization studies, simulation studies, and a combination of the two. Many articles in optimization studies use analytical methods to address the appointment scheduling problem. Although the major benefit of analytical methods is their ability to reach optimal solutions, they may not easily represent all the details and constraints of complex systems. Therefore, many analytical methods simplify the system or relax some of the constraints in order to solve the optimization problem. For instance, the queuing theory is an analytical method that is widely used to address the clinic appointment scheduling problem. Cayirli and Veral (2003) stated that most studies in this domain assumed steady state behavior for the system, which is hardly achievable in healthcare environments. They further added that many optimization methods considered only single-stage systems or made strong assumptions on the distribution of the service time. For instance, special properties of exponential or Erlang distributions used for service times in outpatient appointment scheduling. In addition, Klassen and Yoogalingam (2009) reported that most proposed analytical methods are only valid for problems with a few patients. Begen, Levi, and Queyranne (2012) addressed the problem of appointment scheduling with general discrete probability distributions. However, they considered a single stage facility. Recently, there have been several reports on applications of genetic algorithms in healthcare scheduling at individual departments (see, e.g., Petrovic, Morshed, & Petrovic, 2011).

Another analytical method is mathematical programming (MP) which has been used in patient scheduling and surgery departments. Similar to other analytical methods, MP cannot easily accommodate the complexities and environmental parameters arise in the complex–large systems. A major shortcoming of MP models (except for stochastic programming) is that they are incapable of addressing the stochastic nature of healthcare scheduling problem. Although stochastic programming can address the uncertainty in patient scheduling, most of the models in this area are either overly simplified or analytically intractable, and have been solved using approximation methods (e.g., see Lamiri, Grimaud, & Xie, 2009; Min & Yih, 2010). Another major concern in stochastic programming modeling methods is that the computation time of reaching the optimal solutions is significantly higher than that of deterministic models. In addition, almost none of the studies that consider multi-stage facilities cover patients with different service sequences. The only exceptions are Pham and Klinkert (2008) and Gartner and Kolisch (2014) which address the deterministic version (in terms of processing times) of the problem.

In the context of outpatient appointment scheduling, Fries and Marathe (1981) proposed a dynamic programming method to determine the number of patients to arrive at the beginning of each time block for their appointment scheduling rule. Wang (1999) studied scheduling using non-linear programming for both static and dynamic problems in a clinic. The author assumed that customer service times were independent and identical exponential distributions while customer arrivals were punctual. Denton and Gupta (2003) proposed a stochastic programming model in which appointment times were determined optimally for a fixed appointment sequence.

With respect to MP in surgery scheduling, Hsu, de Matta, and Lee (2003) presented a deterministic two-stage no-wait flow shop model for an ambulatory surgery clinic. The first stage addresses the operating room (OR), surgeons, and scarce resources; the second stage models the post anesthesia unit (PACU). They proposed a heuristic to solve the model with the goal of minimizing the number of PACU nurses and the makespan. Guinet and Chaabane (2003) developed a no-wait flow shop method for inpatient surgery. However, they did not provide any solution method. Ozkarahan (1995) introduced a deterministic mixed integer programming (MIP) model to assign the surgery cases to operating rooms (ORs) with the goal of minimizing the under-time and overtime. In addition, the author developed priority rules to sequence the patients. Sier, Tobin, and Mcgurk (1997) suggested a mixed integer non-liner programming model to assign surgery time blocks to patients. The model considered a penalty function including patient's age and resources such as scarce equipment and ORs. They proposed a simulated annealing approach to solve the model. Pham and Klinkert (2008) proposed a deterministic MIP model based on multi-blocking job shop scheduling problem to minimize criteria such as makespan in surgery-case scheduling. They defined each surgery as a sequence of predetermined jobs. They imposed precedence and priority relations to address the conflict of shared