2000

# Discovering Scheduling Knowledge Through Project Data Mining

Dan Zhu
*Iowa State University*, dzhu@iastate.edu

# Discovering Scheduling Knowledge Through Project Data Mining

Dan Zhu
Department of Logistics, Operations and MIS
Iowa State University
Ames, IA 50011
515-294-5041
515-294-2534 (fax)
e-mail: dzhu@iastate.edu

## Abstract

This project investigates multiple strategies for automatic discovery of scheduling knowledge from projects. In managing large projects with resource and budget constraints, project managers are often overwhelmed with complex project conditions and the large number of available tools. When faced with several alternatives that initially appear equally desirable, making a good judgment can be tedious and often a painful process. The goal of this project is to develop a hybrid system which can induce the relationship between project characteristics and heuristic scheduling performance, and extract meaningful rules to guide managers in mining large project data. Preliminary experiments are conducted and demonstrate promising results. The modular structure of the system will facilitate the dynamic adaptation and learning to accommodate different project conditions. The resulting work will serve as a prototype system for further exploration of intelligent multi-agent architecture.

## Introduction

The goal of this research is to investigate strategies for automatic discovery of scheduling knowledge from project management. In managing large projects with resource and budget constraints, project managers are overwhelmed with complex project conditions and the large number of available solvers. Solver selection for many complex practical applications is a difficult problem due to the availability of a large number of solution procedures, and the heavy requirements on the collection, storage and retrieval of information needed by the solvers. This project proposes to develop a hybrid system to aid the process of heuristic selection, control, and management for the resource-constrained project scheduling problem with cash flows (RCPSPCF). It involves analysis, design, implementation and evaluation of a hybrid system which can induce the relationship between project characteristics and heuristic scheduling performance, and extract meaningful rules to guide managers in mining large project data.

Research on knowledge discovery and data mining has received considerable attention from both academia and industry communities [2]. Machine learning techniques have been used for developing adaptive system in intelligent manufacturing and scheduling [3, 7, 9]. Connectionist models deliver a promising vehicle for exploring the performance of these heuristics with respect to their dependencies on the project parameters. They have been shown to be powerful prediction and classification tools and provide new opportunities for difficult problems that have been traditionally modeled by statistical approaches. However, it is hard to capture what is learned by the neural networks. It is important for management to understand the system. Managers will be reluctant to follow the recommendations if they do not know how the systems arrive at their decisions. This project aims to integrate several methodologies to address the issue of solver selection in order to accomplish scheduling in the resource-constrained project management domain. Learning capabilities will be instilled into this hybrid agent system for developing an intelligent decision support for efficient and effective management. The resulting decision support system will have unique features of being an adaptive and comprehensive tool for project management. The proposed study may also shed light in training novices and unskilled schedulers to handle complex projects.

## Background

The necessity of resolving resource conflicts and cash flow allocations gives rise to resource constrained project scheduling and net present value objective problems. The resource constrained project scheduling problem (RCPSP) is concerned with the scheduling of a project consisting of a number of activities which are linked by precedence relations and multiple resource restrictions. When cash flows exist in the form of expenses for initiating activities and progress payments for completed work, the development of project schedules which maximize the net present value (NPV) of the project is of considerable practical importance. This is a difficult combinatorial optimization problem which precludes the development of optimal schedules [4]. Project management and scheduling deals with tremendous amounts of data and sophisticated interactions with human problem solving. Given the intractable nature of the problem, a number of knowledge sources that are relevant to project scheduling have been identified such as various mathematical models and heuristic procedures.

While many heuristic approaches have been proposed for this problem in the literature, it has been found that no single heuristic can be relied upon to provide the best schedule in all project environments. Different approaches to a problem generally produce distinct algorithms that have significantly different performance characteristics in varying problem environments. In most instances, a single algorithm is chosen to be applied throughout the solution process of a given problem, thus losing the opportunity to exploit the desirable features of other methods. It was observed in earlier studies that heuristic performance varies where there is a change in project parameters. These may include project size parameters, critical path analysis parameters, resource based parameters, cash flow parameters and parameters describing project network shape.

## Proposed Approach and Preliminary Results

Recognizing the need to compare these heuristic procedures and identify the critical project parameters for each, previous research has utilized regression models and neural networks to infer the relationship between heuristic performance and project characteristics [6, 8]. Neural networks offer an alternative which is versatile and may improve the generalization performance. The advantages of the neural networks include their flexibility and tolerance of fuzzy information in generalization power. However, a common pitfall of connectionist models is their black-box nature which lacks explanation. While this may not be critical to the problems in an engineering or financial setting, comprehensible knowledge is especially important to the management. Symbolic rules which are within human comprehension are more convincing and easier to be accepted by project managers. Inductive learning includes learning methods that attempt to induce general concepts from examples. In order to accurately induce the hidden relationships in projects conditions and solvers' performance, several approaches will be investigated. One of such is called RIPPER.

RIPPER [1] is a rule learning algorithm improved on the basis of another rule induction method, Incremental Reduced Error Pruning (IREP). Like a standard separate-and-conquer algorithm, IREP builds up a rule by adding one rule at a time. After a rule is found, all examples covered by this rule are deleted. The process is repeated until there are no positive examples, or there is an unacceptable error rate created by IREP. Although it has more promising computational efficiency than that of C4.5 [5], especially for noisy data, its generalization performance needs to be improved. Cohen (1995) made several modifications to this algorithm and considerably improved its generalization performance. The three modifications are as follows: 1) Replace the rule-value metric with a new one that has more intuitively satisfying behavior. 2) Implement a new heuristic condition for the learning algorithm to stop adding rules to the rule set. After each rule is added, the total "description length" of the rule set and the examples are computed. When this length is more than d bits longer than the smallest description length, the learning system stops adding new rules. 3) An optimizing method is used. Besides the original rule, there are two alternative rules, the revised rule and the replacement rule, constructed for each. A heuristic is used to judge which rule should be added. Through these modifications, a new learning algorithm called RIPPER was developed. The basis strategy for RIPPER to find a rule set is to first use IREP to find an initial model, and then to improve that model, with the optimization procedure described above. Thereby, RIPPER gets better performance than that of IREP, and it is competitive with C4.5 rules in error rates, but much more efficient on large and noisy data sets.

In this project, an extensive examination of all the

parameters that have been identified in the literature for the resource-constrained project scheduling problem were conducted. To analyze the effect of these problem parameters on heuristic performance, numerous heuristic rules identified from literature were evaluated. Six parameters were obtained through a reduct from about 30 project characteristics identified from the literature. Correspondingly, a category of four clusters of heuristics were included in investigating heuristic performance, with rules being grouped based on similar type of information such as dual price and opportunity cost. The final database consists of close to 1500 projects and their corresponding NPVs obtained using different heuristics. This dataset was randomly partitioned into several parts for training, testing and validation. A 3-fold cross-validation have been run using RIPPER with about 70% for training and 30% for testing. The average prediction accuracy is 68% which is slightly better than that from the neural network studies [8]. Additionally, four decision rules have been obtained. The experiments were conducted on a HP workstation and the running time was fast (in terms of seconds).

## Conclusion and Discussion

When faced with several alternatives that initially appear equally desirable, making a good judgment can be tedious and often a painful process. This paper introduces a new technology to address the problem of solver selection in an attempt to accomplish the task of knowledge discovery in project management and scheduling. Initial experiments and results deomonstrate promising performance. Many important issues such as problem representations, data preprocessing and system integration will be explored. The modular structure of the system will facilitate the dynamic adaptation and learn to accommodate different project conditions.

Successful data mining in project management requires extensive effort and a diverse set of skills. This research is intended to support the complex decision making process in project scheduling and management. Using such a system, project managers can make timely decisions by examining critical project characteristics and learning to predict the effectiveness of various scheduling rules under different situations. Furthermore, project managers may use this system as a guide to dynamically switch heuristic rules as the scheduling progresses.

## Acknowledgment

This research is partially supported by NSF and a research development grant from Iowa State University. I would like to thank Xiaoning Zhang for the help with the experiments.

## References

[1] William W. Cohen. Fast effective rule induction. In *the Proceedings of the Twelfth International Conference on Machine Learning..*, 1996.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), Nov. 1996.

[3] M.S. Fox. *Constraint-Directed Search: A Case Study in Job Shop Scheduling.* Morgan Kaufmann, Palo Alto, CA, 1987.

[4] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman and Co., 1979.

[5] Quinlan J.R. *C4.5 Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1992.

[6] J.H. Patterson. Project scheduling: The effects of problem structure on heuristic performance. *Naval Research Logistics Quarterly*, 23(1):95–122, 1976.

[7] M. Shaw, S. Park, and N. Raman. Intelligent scheduling with machine learning capabilities: The induction of scheduling knowledge. *IEEE Transactions on Engineering Management*, (2), 1992.

[8] D. Zhu and R. Padman. Connectionist approaches for solver selection in constrained project scheduling. *Annals of Operations Research*, 72(2):302–317, 1997.

[9] D. Zhu, M. Prietula, and W. Hsu. When processes learn: Steps toward crafting an intelligent organization. *Information Systems Research*, 93(3):302–317, 1997.