The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0972-7981.htm

Implementation of supervised statistical data mining algorithm for single machine scheduling

S. Premalatha
Department of Computer Applications, J.J. College of Engineering and Technology, Tiruchirapalli, India, and
N. Baskar
Department of Mechanical Engineering, M.A.M College of Engineering, Tiruchirapalli, India

Abstract

Purpose – Machine scheduling plays an important role in most manufacturing industries and has received a great amount of attention from operation researchers. Production scheduling is concerned with the allocation of resources and the sequencing of tasks to produce goods and services. Dispatching rules help in the identification of efficient or optimized scheduling sequences. The purpose of this paper is to consider a data mining-based approach to discover previously unknown priority dispatching rules for the single machine scheduling problem.

Design/methodology/approach – In this work, the supervised statistical data mining algorithm, namely Bayesian, is implemented for the single machine scheduling problem. Data mining techniques are used to find hidden patterns and rules through large amounts of structured or unstructured data. The constructed training set is analyzed using Bayesian method and an efficient production schedule is proposed for machine scheduling.

Findings – After integration of naive Bayesian classification, the proposed methodology suggests an optimized scheduling sequence.

Originality/value – This paper analyzes the progressive results of a supervised learning algorithm tested with the production data along with a few of the system attributes. The data are collected from the literature and converted into the training data set suitable for implementation. The supervised data mining algorithm has not previously been explored in production scheduling.

Keywords Programming and algorithm theory, Production scheduling, Data mining, Dispatching rule, Learning algorithm, System attributes

Paper type Research paper

1. Introduction

Data mining is often referred to as knowledge discovery in databases. Data mining has made broad and significant progress since its early beginning in the 1980s. Recently data mining techniques (Kriegel *et al.*, 2007) have been applied to various operations researches, engineering and medical applications. A data mining-based approach to discover previously unknown priority dispatching rules for a single-machine scheduling problem is considered. This approach involves the generation of a flat production data file from the problem data and the output is an efficient scheduling rule using a supervised statistical data mining algorithm. Scheduling mechanisms are typically implemented by means of a knowledge base obtained through a machine learning-based dynamic dispatching mechanism on the basis of the current system status. A data mining-based scheduling framework is presented and illustrated using a simple example.

Journal of Advances in Management Research
Vol. 9 No. 2, 2012
pp. 170-177

0972-7981
DOI 10.1108/09727981211271913

The scheduling can be approached in two ways: static scheduling consists of fixed sets of n-jobs to be run. It is performed using deterministic processing times and stochastic processing times. Another is dynamic scheduling, where new jobs are continually being added over time. Processing times for these jobs can be either deterministic or stochastic. The single-machine scheduling problems consist of independent jobs, each with a single operation. The two key problems in production scheduling according to Wight (1984) are priorities and capacity. Scheduling problems often occur when different jobs or orders have to be processed on one or multiple machines. The Bayesian approach is concerned with processing probabilistic information represented in the form of prior probabilities and conditional probabilities.

2. Literature review

Optimization of scheduling problems is one of the key elements for increasing production rate in manufacturing industry. The process planners are followed by the recommendation of process manager and the manager schedules the job based on the need of the customers. Most of the researchers scheduled the job as per the objective function of the problem. Li and Olafsson (2005) proposed a methodology for generating scheduling rules using a data-driven approach. He proposed a model with a Decision Tree for dispatching rule selection considering processing time and release time. Kumar and Rao (2009) proposed the use of data mining algorithms for the extraction of knowledge from a large set of flow shop schedules. They delivered an approximate method to resolve a multi-product batch flow shop schedule. A data mining tool used by them for the study was the Decision Tree. Yeou-Ren Shiue (2009) proposed a two-level Decision Tree learning approach for dynamic dispatching rule selection mechanism under various performance criteria over a long period. The various inter-arrival time effects are not explored. Ei-Bouri and Shah (2006) investigated the application of a neural network for selecting, from a set of available dispatching rules. Nazif and Lee (2009) developed an optimized crossover genetic algorithm to effectively solve the single-machine scheduling problem which minimizes the total weighted completion time of the jobs in the presence of the sequence independent family setup times.

Subramanian *et al.* (2000) proposed that the significant improvements to the scheduling performance are realized when the machine selection rules are used. Quinlan (1986) demonstrated the technology for building Decision Trees. Janiak and Krysiak (2007) proposed exact polynomial time algorithms for the NP-hard case of single processor scheduling. They proposed a method to minimize the weighted number of late jobs. Dirk Biskup (2008) reviewed scheduling with learning effects for efficient scheduling decisions. Xue Huang *et al.* (2010) discussed a single-machine scheduling problem with deteriorating jobs in which the due dates are determined by the equal slack method. Also they proved that the problem can be solved in polynomial time. Baptiste *et al.* (2010) proposed an integer linear programming formulation to minimize the weighted number of tardy jobs on a single machine when both due dates and deadlines are specified for jobs. Olafsson and Li (2010) proposed a Decision Tree learning from the scheduling data by applying a classification method. Yang *et al.* (2010) proposed an important constraint satisfaction adaptive neural network for job-shop scheduling problems. Harding *et al.* (2006) reviewed the applications of daa mining in manufacturing engineering, in particular production processes, operations, fault detection, maintenance, decision support, and product quality improvement. Many researchers found a significant improvement in the scheduling problem by using

optimization techniques. This work is mainly concentrated on finding the best sequence for combining both optimization and data mining algorithms for machine scheduling.

3. Problem description

Scheduling is the allocation of resources over time to perform a collection of tasks. The scheduling can be applied for both static scheduling problems and dynamic scheduling problems. The problem that is considered for the study consists of n jobs, each with single operation. Set-up time of each of the jobs is independent of its position in the sequence of jobs. Each job is processed until its completion without pre-emption. The dispatching rules considered for the study are shortest processing time (SPT), weighted shortest processing time (WSPT). The mean flow time is calculated as:

$$\overline{F} = \frac{1}{n}\sum_{j=1}^{n} F_j$$

where $F_j = C_j - r_j$, C_j is the completion time, r_j the ready time and $F_j = C_j$ if $r_j = 0$.

Basic data necessary to describe jobs in a deterministic single-machine scheduling problem are: *Processing time (t_j)*: the time required to process job j. It includes both the actual processing time and set-up time:

(1) *Ready time (r_j)*: the time at which job j is available for processing. It is the difference between the arrival time of that job and the time at which that job is taken for processing.

(2) *Due date (d_j):* the time at which the processing of the job j is to be completed.

(3) *Completion time (C_j):* the time at which the job j is actually completed in sequence.

(4) *Flow time (F_j):* the amount of time that job spends in the system. It is the difference between the completion time and the ready time of the job j. $F_j = C_j - r_j = C_j$ if $r_j = 0$.

(5) *Lateness (L_j):* the amount of time by which the completion time of job differs from its due date.

(6) *Tardiness (T_j):* the lateness of job j if it fails to meet its due date; otherwise it is zero:

$$T_j = \max\{0, C_j - d_j\} = \max\{0, L_j\}$$

where $T_j = C_j - d_j$ if $C_j > d_j = 0$ otherwise.

The performance is measured using the above said data. Among the various supervised learning algorithms, Bayesian method is applied to ensure the optimized scheduling sequence. The training data set is collected from the literature.

4. Methodology

The data mining algorithms are the mechanism that creates a data mining model. To create a model, an algorithm first analyzes a set of data developed for specific

patterns and trends. The algorithm uses the results of the analysis to define the parameters of the mining model. These parameters are then applied across the entire set to extract actionable patterns and detailed statistics. Among various supervised learning algorithms like Bayesian method, regression, Decision Trees, rule algorithms, hybrid algorithms and neural networks, the Bayesian method is adopted for the study.

4.1 Supervised statistical data mining algorithm

The statistical data mining algorithms namely Bayesian method is used to propose an optimal dispatching sequence for the single-machine scheduling problem.

4.1.1 Naïve Bayesian classification algorithm (Han and Kamber, 2006). Bayes' theorem is named after Thomas Bayes', who did early work in probability and decision theory during the eighteenth century. Naive Bayes' (Wu *et al.*, 2008) classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naïve Bayes' models uses the method of maximum likelihood; in other words, one can work with the naïve Bayes' model without believing in Bayesian probability or using any Bayesian methods. Naive Bayesian classifiers have worked quite well in many complex real-world situations. Statistical processing based on the Bayes' decision theory is a fundamental technique for pattern recognition and classification. It is based on the assumption that the classifications of patterns (the decision problem) are expressed in probabilistic terms. The statistical characteristics of patterns are expressed as known probability values that describe the random nature of patterns and their features. These probabilistic characteristics are mostly concerned with a priori probabilities and conditional probability densities of patterns and classes. The Bayes' decision theory provides a framework for statistical methods for classifying the patterns into classes based on probabilities of patterns and their features. Let X is a data tuple. In Bayesian terms, X is considered as the evidence. It is described the measurements made on a set of n attributes. Let H be the hypothesis, such as that the data tuple X belongs to a specified class C.

Bayes' theorem is:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

where $P(H|X)$ is the posterior probability, of H contained on X and $P(X|H)$ the posterior probability, of X contained on H. $P(X)$ is the prior probability of X.

The Bayesian classification for single-machine scheduling problem is shown in Figure 1.

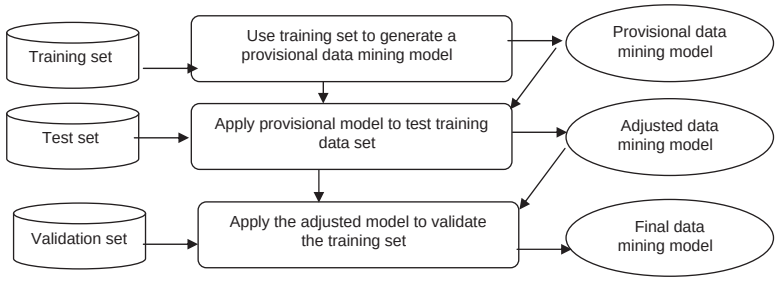

Figure 1.
Bayesian classification for single machine scheduling

5. Implementation

The dispatching list proposed by Olafsson and Li (2010) is considered in this work. Table I shows the dispatching list of the proposed problem.

Algorithm:

Input: Scheduling sequence generated by SPT/WSPT dispatching rules.

Step 1: Collect the attributes to be included and generate training samples.
Step 2: Build knowledge database file from the production data.
Step 3: Find the probability of scheduling the job as follows:

3.1 Identify the release time probability of Job 1 than Job 2 (Job 1 releases earlier).
3.2 Identify the processing time probability of job1 than job2 (Job 1's processing time is lower).

Step 4: Apply naïve Bayesian classification and find the probability (yes and no).
Step 5: Repeat Steps 3-4.
Step 6: Compare the results with the flow time and dispatching sequence given in the training set (SPT and WSPT).
Step 7: Propose the optimized sequence.

6. Results and discussion

Data mining and knowledge discovery are the emerging areas of research and applications that draw on machine learning and statistical methods to learn previously unknown and useful knowledge from the large database. The study mainly focussed on the attributes like release time, processing time and weight for classification of the data. The classification of data is made with the training data file. The maximum number of training set can be found using the following formula:

$$\sum_{J=1}^{n-1} (n-j)$$

Total number of training sets for the sample is 10 and the number of subsets can be 2^{10}. The jobs to be released first and subsequently are chosen. The comparison is done with the following dispatching rules: SPT and WSPT.

The dispatching criteria is used to schedule Job 1 first:

(1) if Job 1 releases earlier and processing time is lower than Job 2 then Job1 is scheduled first (yes);

(2) if Job 1 releases or processing time is not lower than Job 2 then Job 1 is scheduled first (yes); and

Table I. Dispatching list

Job ID	Release time	Due date	Processing time	Weight
1	2	24	7	5
2	0	7	3	3
3	15	10	4	1
4	5	36	18	3
5	3	25	6	3

(3) if Job 1 is not released first and the processing time of Job 1 is not lower, then Job 2 is scheduled first (no) (see Table II).

6.1 Performance evaluation

By applying naïve Bayesian theorem it is found that the probability to accept the SPT dispatching schedule is higher than WSPT and other alternative dispatching schedules. It is also found that the earlier release time jobs if scheduled first will also yield good probability. Sample probability values obtained for the Schedule 1: Job 2 – Job 3 – Job 5 – Job 1 – Job 4 is 0.5 and for the Schedule 2: Job 1 – Job 2 – Job 5 – Job 3 – Job 4 is 0.432 and for the Schedule 3: Job 1 – Job 2 – Job 3 – Job 4 – Job 5 is 0.144.

Hence it is clear that to state that the Naïve Bayesian theorem evaluated Schedule 1 as the optimized schedule. Here the study focussed mainly on Bayes' theorem in single-machine scheduling problem. The same study could be extended to other data mining algorithms.

7. Managerial implications of research

Based on the reviews conducted it is clear that the scheduling problem was handled by evolutionary algorithms and many conventional algorithms. And the Decision Tree method was used. But, the proposed methodology aids the manufacturing industries in the following ways:

- A new methodology by integrating Naïve Bayesian classification algorithm is used to deduce production schedule.
- The tool produces effective schedules in less amount of time. The optimized schedules are always predicted.
- The generated schedule can be stored along with the input data set in a data warehouse to infer knowledge if such job patterns occur in near future.

8. Conclusion

Recently data mining has impacted many industries but in production scheduling and manufacturing industry it has received less attention. This work explains how data

Job 1	RT1	PT1	Job 2	RT2	PT2	Job 1 releases earlier	Job 1 processing time is lower	Job 1 scheduled first
2	0	3	3	15	4	Yes	Yes	Yes
2	0	3	5	3	6	Yes	Yes	Yes
2	0	3	1	2	7	Yes	Yes	Yes
2	0	3	4	5	18	Yes	Yes	Yes
3	15	4	5	3	6	No	Yes	Yes
3	15	4	1	2	7	No	Yes	Yes
3	15	4	4	5	18	No	Yes	Yes
5	3	6	1	2	7	No	Yes	Yes
5	3	6	4	5	18	No	Yes	Yes
1	2	7	4	5	18	Yes	Yes	Yes

Notes: RT, release time; PT, processing time

Table II.
Training data set for the shortest processing time dispatching list

mining algorithms can be implemented to single-machine scheduling for the given objective function. Even though scheduling problems can be deterministic, it becomes NP-hard as the uncertainty situations occur. The various forms of uncertainty situations are the job activities that take more or less time than the originally estimated time, resources or machines that may become unavailable, material arrival behind the schedule time, change in due dates and ready times, necessity to incorporate new activities and to drop certain activities, delays due to natural calamities, etc. This paper analyses the training data set obtained for various dispatching rules with the help of supervised statistical algorithm. The same could be analyzed by increasing the number of attributes. Also significant improvements can be made and new data mining techniques can be applied to extend the knowledge of single machine, job-shop and flow-shop problems.

References

Baptiste, P., Croce, F.D., Grosso, A. and Kindt, V.T. (2010), "Sequencing a single machine with due dates and deadlines: an ILP-based approach to solve very large instances", *Journal of Scheduling*, Vol. 13 No. 1, pp. 39-47.

Biskup, D. (2008), "A state-of-the-art review on scheduling with learning effects", *European Journal of Operational Research*, Vol. 188, pp. 315-29.

Ei-Bouri, A. and Shah, P. (2006), "A neural network for dispatching rule selection in a job shop", *International Journal of Advanced Manufacturing Technology*, Vol. 31 Nos 3-4, pp. 342-9.

Han, J. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Haryana.

Harding, J.A., Shahbaz, M., Srinivas and Kusiak, A. (2006), "Data mining in manufacturing: a review", *Journal of Manufacturing Science and Engineering*, Vol. 128 No. 4, pp. 969-76.

Huang, X., Wang, J.-B. and Wang, X.-R. (2010), "A generalization for single-machine scheduling with deteriorating jobs to minimize earliness penalties", *International Journal of Advanced Manufacturing Technology*, Vol. 47 Nos 9-12, pp. 1225-30.

Janiak, A. and Krysiak, T. (2007), "Single processor scheduling with job values depending on their completion times", *Journal of Scheduling*, Vol. 10 No. 2, pp. 120-38.

Kriegel, H.-P., Borgwardt, K.M., Kroger, P., Pryakhin, A., Matthias, S. and Zimek, A. (2007), "Future trends in data mining", *Journal of Data Mining and Knowledge Discovery*, Vol. 15, pp. 87-97.

Kumar, S. and Rao, C.S.P. (2009), "Application of ant colony, genetic algorithm and data mining-based techniques for scheduling", *Robotics and Computer-Integrated Manufacturing*, Vol. 25 No. 6, pp. 901-8.

Li, X. and Olafsson, S. (2005), "Discovering dispatching rules using data mining", *Journal of Scheduling*, Vol. 8 No. 6, pp. 515-27.

Nazif, H. and Lee, L.S. (2009), "A genetic algorithm on single machine scheduling problem to minimize total weighted completion time", *European Journal of Scientific Research*, Vol. 35 No. 3, pp. 444-52.

Olafsson, S. and Li, X. (2010), "Learning effective new single machine dispatching rules from optimal scheduling data", *International Journal of Production Economics*, Vol. 128 No. 1, pp. 118-26.

Quinlan, J.R. (1986), "Induction of decision trees", *Journal of Machine Learning*, Vol. 1 No. 11, pp. 81-106.

Shiue, Y.-R. (2009), "Development of two-level decision tree-based real-time scheduling system under product mix variety environment", *Robotics and Computer-Integrated Manufacturing*, Vol. 25 Nos 4-5, pp. 709-20.

Subramanian, V., Lee, G.K., Ramesh, T., Hong, G.S. and Wong, Y.S. (2000), "Machine selection rules in a dynamic job shop", *International Journal of Advanced Manufacturing Technology*, Vol. 16 No. 12, pp. 902-8.

Wight, O.W. (1984), *Production and Inventory Management in the Computer Age*, Van Nostrand Reinhold Company Inc, New York, NY.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J. and Steinberg, D. (2008), "Top 10 algorithms in data mining", *Knowledge Information System*, Vol. 14 No. 1, pp. 1-37.

Yang, S., Wang, D., Chai, T. and Kendall, G. (2010), "An improved constraint satisfaction adaptive neural network for job-shop scheduling", *Journal of Scheduling*, Vol. 13 No. 1, pp. 17-38.

Further reading

Jiang, X. and Cooper, G.F. (2010), "A real-time temporal Bayesian architecture for event surveillance and its application to patient-specific multiple disease outbreak detection", *Journal of Data Mining and Knowledge Discovery*, Vol. 20 No. 3, pp. 328-60.

Corresponding author

S. Premalatha can be contacted at: lathachand78@gmail.com