A Hybrid Methodology Based on Machine Learning for a Supply Chain Optimization Problem

Duy Nguyen Duc and Narameth Nananukul*
School of Management Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani 12120, Thailand
*Corresponding author email: narameth@siit.tu.ac.th

Abstract. This paper presents an advanced methodology that integrates a machine learning methodology into an optimization process. The framework of an interactive machine learning algorithm was developed to meet the challenges in solving large-scale optimization problems. An artificial neural network (ANN) is used with the knowledge gained from solving previous problems with different scenarios to define a good starting point for a solution searching process. By using an initial solution, known as “warm start”, the search space can be reduced to get more opportunity to find an optimal solution. The applicability of the proposed method was evaluated by using it to determine the optimal facility locations for a biomass supply chain problem using a real case study from Central Vietnam. The supply chain planning model is based on an optimization model, where the goal is to maximize the benefits from meeting the electricity demand minus the total cost from facility cost, penalty cost from lost demand, and operational costs form the supply chain. The structure of the ANN, the number of intermediate layers and the number of processing nodes, was determined by comparing the accuracy from different configurations. The ANN with two intermediate layers possesses the best performance from the training and testing datasets. The proposed model succeeded in predicting the facility location with more than 98% prediction accuracy. The results from our framework provide optimal solutions while saving runtime.

Keywords: Machine learning; Optimization; Mixed integer programming; Supply chain planning; Artificial neural network (ANN); Relaxation Induced Neighbourhood Search (RINS).

1. Introduction
In general, the supply chain optimization process can be used to determine the optimal operational plans for the manufacturing and logistics functions within a supply chain. Due to its practical application, it has become an interesting field for academia and practitioners. Data mining is a technique that can be used to extract knowledge from vast datasets. Computer system enables the discovery of hidden information in datasets. When applying data mining with optimization, once the hidden information from the data is extracted, the search space of an optimization problem can be reduced. As a result, an optimal solution or good quality solution can be determined within a limited time and computer capacity. To demonstrate the effectiveness of this approach, a machine learning methodology for optimization is applied to a constrained optimization problem, and the results are discussed.

Vietnam, as an agricultural Asian country, has high potential for biomass energy utilization. The demand for biomass energy in Vietnam can be found in [1]. In this study, a framework that integrates a machine learning methodology with an optimization process for a biomass supply chain is introduced. The proposed model is based on [1] where large scale problems were solved by a hybrid...
simulation-based optimization approach. The uncertainty from the demand and supplier capacity were captured by a simulation model, which reduces the complexity of the original optimization problem. In this research, the search space of an optimization problem is reduced by using the cumulative knowledge gained from solving past optimization problems with different scenarios, this approach is implemented by using an artificial neural network (ANN) model. An initial solution from the machine learning model was fed to a mixed integer programming (MIP) solver, CPLEX. By combining a “warm start” or “advantage start” with Relaxation Induced Neighbourhood Search (RINS) function from CPLEX, the optimal solution can be found with practical time. Data related to biomass supply chain from Vietnam were considered for planning biomass plant locations. This is considered a main contribution of this research.

The content of the paper is described as follows. A literature review of the stochastic model for supply chain management, machine learning with optimization (warm start), and RINS for solving MIP is provided in section 2. Section 3 presents a data encoding scheme for the machine learning model. A case study for a biomass supply chain planning is used to illustrate the methodology with discussion in Section 4. Section 5 presents the conclusion and general guidelines for future work.

2. Literature Review

The application of machine learning to assist supply chain planning was successfully implemented by many researchers. Silva et al. [2] used an ANN to provide control/visibility over a supply chain in a planning time frame. Their model acts pro-actively to avoid disruptions by predicting the stated upcoming orders and foreseeing which entity will reach its re-order point, and purchase new orders. Lima-Junior and Carpinetti [3] ANNs to estimate lagging metrics which are dependent on leading metrics based on causal relationships. An ANN can deal with the challenges of adjusting the performance prediction by using the average of past performance data. A case study in Nigeria solar energy was proposed by Ozoegwu [4]. The solar energy available at selected sites was predicted by a hybrid ANN model. Cai et al [5] constructed a risk evaluation model by using a multilayer feedforward neural network. Akbarian-Saravi et al. [6] used an ANN to predict the bioethanol demand for biomass supply chain model which was developed for a strategic and tactical design.

Considering the usage of machine learning to guide solution searching in a search space of an optimization problem, Chen and Huang [7] introduced a hybrid global search, where a machine learning model was used to define a search space. Then, an evolution strategy was used to find a nearly global optimal solution. An unsupervised learning method was applied to identify potential regions within a large search space by Ghiani et al. [8]. Then, four types of neighbourhood search methods were applied and compared based on the results. A multi-objective memetic algorithm that is based on a machine learning methodology was developed by Wang and Tang [9] for a flow shop scheduling problem with multiple objective. Representative solutions for improvement were selected based on a result from a clustering process. Then, a solution within the promising regions was selected by using a statistical learning method. Bagloee et al. [10] investigated a combination of machine learning and optimization for a contraflow design problem, which is known as NP-hard. The algorithm was outstanding when compared with other metaheuristic algorithms, using experimental results from two real cases based on different network sizes.

Providing a hint to help an optimization solver such as CPLEX to search for an optimal solution for a MIP can be done by setting a MIP start, or an advanced start [11]. Pour et al. [12] used an initial solution from a constraint programming model to feed an optimization model as a warm start solution. Their approach can reduce the size of a search space of a scheduling problem and was able to solve problems up to an 8-week period, while a traditional mixed-integer programming method can only solve up to a 2-week planning period. Li et al. [13] introduced a three-level heuristic for a production-routing large-scale problem. Three levels were defined as: initial solution generation, infeasibility repair, and incumbent solution improvement. Experimental results indicated that a three-level heuristic can efficiently solve large-sized problems.

Instead of starting with any feasible solution, a “warm start” can provide a good initial solution for future improvement. After this process, RINS heuristic can be applied to determine an optimal solution [14]. D’Andreagiovanni et al. [15] solved a robust multi-period network design problem by
developing a fast primal heuristic that is based on a combination of an ant colony heuristic method and a neighbourhood search. Their approach was illustrated and verified by 30 datasets. A cooperation between a genetic algorithm and RINS intended for a local branching (LB) was introduced by [16]. The results showed that in many cases, RINS provided good result. The following provides a summary of contribution from this research: first, the insight gained from a biomass supply chain case study used in this research can be generalized to the biomass supply chain planning from other cases. Second, this research proposes an integration of machine learning to an optimization process for generating a solution for a biomass supply chain problem which increases the computational efficiency.

3. Dataset for an Artificial Neural Network Model
In this section, the description of data used for building an ANN is provided. The configuration of the model is provided in the following sections.

3.1. Data Preparation
Based on the relationship between the facility location and input data, an ANN model was developed as shown in figure 1. Factors such as electricity demand, purchasing cost, holding cost, transportation cost, and supplier’s capacity were defined as input factors. The target output is factory location. The data was first pre-processed in order to reduce the redundancy of data in the dataset. All elements from both the input and target that were unchanged or equal to zero in all samples were considered redundant and can be omitted. Because the considered biomass supply chain model is large and contains many types of data with different ranges; therefore, a data normalization was applied to standardize variables with different units. A min-max normalization method was chosen to normalize input data. Then, a chi-square feature selection method [17] and principal component analysis [18] were applied to eliminate nonessential data which improves the generalization ability of the model [18]. Then, training and testing datasets were defined for the ANN model. The training data were used to develop an ANN model for factory location prediction, while the testing data were used to validate the performance of the model. Seventy percent of the data was for training the model, while the remaining data was for testing the model. The accuracy from configurations with different number of intermediate layers and number of nodes was used to determine the best configuration.

3.2. “Warm start” and RINS
In this section, an approach that integrates “Warm start” to RINS is described. When using a “warm start”, the result from the ANN model, factory location, was used to set an initial solution of the biomass supply chain model. This can be implemented by setting a parameter in a MST file, as shown in figure 2. Then, the initial solution will be fed to a CPLEX MIP solver (RINS) where the solution will be improved by the solver. For RINS approach, the initial solution will be improved by a heuristic algorithm which can generate solution with practical runtime. A report on performance of this approach is provided in section 4.

4. A Case Study, ANN Model, Computational Results, and Discussion
An optimization model based on a biomass supply chain optimization model [1] was used to generate input and output datasets.
4.1. A Case Study
A set of biomass supply chain data from a region of Vietnam was used in the case study. The details of a data collection method can be found in [1] and [19]. IBM ILOG CPLEX, an optimization solver, was used to implement optimization models. The ANN model was developed by using an ANN toolbox in Matlab 2019a. The processing computer was equipped with an Intel Core i7 CPU, and 8.0 GB RAM. The operating system is a 64-bit Windows 10. In this research, we focus on 12 agricultural residues that can be used for biomass energy production. The problem considers 12 periods per year, 19 suppliers, and 50 biomass types. The input data for the optimization model were generated randomly until 1500 samples were obtained. Optimal solutions of the optimization model were also determined by using CPLEX. A set of random data was generated based on the original data, where the applied distributions are described as follows: Demand follows a random uniform [0.5, 1.5] distribution; purchasing cost, holding cost, and transportation cost are based on a random uniform [0.5, 2] distribution; supplier capacity follows a random uniform [0, 2] distribution.

4.2. ANN Model
A machine learning model and how the data were prepared are presented in this section. Based on an analysis of all possible factory locations from the output of the optimization model, the factory locations were grouped into 7 patterns as shown in table 1. An encoding where numbers from 1 to 7 are used to represent the patterns is not suitable based on our experiment. To overcome this problem, we use an “One Hot Encoder” encoding scheme as shown in table 1.

| Output Pattern | Output code | Output Pattern | Output code |
|----------------|-------------|----------------|-------------|
| Y3 and Y49     | 1 1 0 0 0 0 0 0 | Y3 and Y16    | 5 0 0 0 0 1 0 0 |
| Y3 and Y35     | 2 0 1 0 0 0 0 0 | Y35 and Y49   | 6 0 0 0 0 0 1 0 |
| Y3 and Y5      | 3 0 0 1 0 0 0 0 | Y1 and Y2     | 7 0 0 0 0 0 0 1 |
| Y49 and Y50    | 4 0 0 0 1 0 0 0 |

A set of data that includes demand, purchasing cost, holding cost, transportation cost, and supplier’s capacity was used as input data, where there were 1543 factors. To reduce redundant data from the dataset, all factors that have no change in value in all samples were omitted because they do not have any effect on the output. We used a min-max normalization [0 -1] method to standardize the unit of all factors, where each factor was normalized separately. From all factors, the factors were screened by using a chi-square test with p-value = 0.05, this reduced the number of factors to 140. Then, a principal component analysis (PCA) was used for further dimension reduction, this resulted in 32 factors.

In general, the performance of an ANN depends on its configuration which is defined by the number of intermediate layers and the number of nodes. In this study, the best configuration of an ANN was determined by using accuracy as a criterion. The one with the highest accuracy was selected. The result is shown in figure 3. The training stage was limited to 1000 iterations. The accuracy was measured by using MSE (Mean Squared Error). The training function for ANN’s input layer was based on a “logsig” function, while for output layer, it was based on a “purelin” function. A training function based on “trainbr” was used for the hidden layer.

Samples were divided into 2 disjoint sets: training dataset (70%) and validation-testing dataset (30%). The testing data set was not used for training the network but has the same range as the training data. This enables the ANN to generalize the learned pattern to provide results for new input datasets. The experimental results were summarized in table 2. Based on the results, the accuracy from the ANN with one hidden layer is not acceptable, where the test accuracy is below 85%. The results from the models using 2 hidden layers shows that using too many numbers of neurons will cause an overfitting...
problem, where the accuracy from the training process (98.5%) is higher than the one from the testing process (96.2%).

![Figure 3. The best configuration of ANN.](image)

### Table 2. Different configurations of ANN with accuracy

| ANN | Accuracy | MSE | ANN | Accuracy | MSE | ANN | Accuracy | MSE |
|-----|----------|-----|-----|----------|-----|-----|----------|-----|
| [32 32 7] | 89% | 70% | 0.0394 | [32 32 10] | 89.0% | 97.2% | 0.0015 | [32 32 20] | 98.5% | 96.2% | 0.0009 |
| | | | | | | | | |
| [32 39 7] | 92% | 80.2% | 0.017 | [32 39 20] | 98.5% | 96.2% | 0.0009 | [32 39 7] | 7 |

### 4.3. Results from using “Warm start” and RINS, and Discussion

The result from the trained ANN model, which is the factory location, is improved by using “Warm start” and RINS. The result is used as an initial solution of a “Warm start” function, and then is improved by RINS which is a heuristic algorithm. The quality of the solutions obtained from the proposed approach was compared with the solutions from a standard MIP solver from CPLEX as shown in Table 3. Table 3 shows the runtime and solutions using the default CPLEX setting and RINS. The optimal solutions obtained from both methods confirms that the factory location decisions are the same. This shows that the ANN can be used to determine a good incumbent solution, and RINS can improve the solution to optimality with practical runtime. The results were benchmarked with the results from another relevant research. For example, a bioethanol supply chain in Brazil was considered by Kostin et al. [20] where they constructed a model that considers multiple production technologies from different production facilities, and a transportation model. As shown in the optimal solution from the model, most facilities use a large amount of raw material, which is mainly sugarcane supplied by suppliers located close to the sites. Likewise, in this research, due to a large fixed cost from plant, only a small number of plants is chosen (small regions use a single plant, while large regions require two plants). The selected plants are located near suppliers with abundant amount of biomass.

### Table 3. Solution comparison between MIP solver and RINS (Factory locations and Runtime)

| Data sets 1 | MIP solver | RINS |
|-------------|------------|------|
| Y3 & Y49 - 1069 s | Y3 & Y49 - 498 s |

| Data sets 2 | MIP solver | RINS |
|-------------|------------|------|
| Y3 & Y35 - 734 s | Y3 & Y35 - 580 s |
References

[1] Duc D N and Nananukul N 2019 Advanced Methodologies for Biomass Supply Chain Planning Processes 7 659
[2] Silva N, Ferreira L M D, Silva C, Magalhães V and Neto P 2017 Improving supply chain visibility with artificial neural networks Procedia Manufacturing 11 2083-90
[3] Lima-Junior F R and Carpinetti L C R 2019 Predicting supply chain performance based on SCOR® metrics and multilayer perceptron neural networks Int. J. of Production Economics 212 19-38
[4] Ozoegwu C G 2019 Artificial neural network forecast of monthly mean daily global solar radiation of selected locations based on time series and month number J. of cleaner production 216 1-13
[5] Cai X, Qian Y, Bai Q and Liu W 2019 Exploration on the financing risks of enterprise supply chain using Back Propagation neural network J. of Computational and Applied Mathematics 112457
[6] Akbarian-Saravi N, Mobini M and Rabbani M 2019 Development of a comprehensive decision support tool for strategic and tactical planning of sustainable bioethanol supply chain: Real case study, discussions and policy implications J. of Cleaner Production 118871
[7] Chen T and Huang J 2013 Application of data mining in a global optimization algorithm Advances in Engineering Software 66 24-33
[8] Ghiani G, Laporte G and Manni E 2015 Model-based automatic neighborhood design by unsupervised learning Computers & Operations Research 54 108-16
[9] Wang X and Tang L 2017 A machine-learning based memetic algorithm for the multi-objective permutation flowshop scheduling problem Computers & Operations Research 79 60-77
[10] Bagloee S A, Johansson K H and Asadi M 2019 A hybrid machine-learning and optimization method for contraflow design in post-disaster cases and traffic management scenarios Expert Systems with Applications 124 67-81
[11] Gondzio J 1998 Warm start of the primal-dual method applied in the cutting-plane scheme Mathematical Programming 83 125-43
[12] Poudel S, Marufuzzaman M, Quddus M, Chowdhury S, Bian L and Smith B 2018 Designing a reliable and congested multi-modal facility location problem for biofuel supply chain network Energies 11 1682
[13] Li Y, Chu F, Chu C and Zhu Z 2019 An efficient three-level heuristic for the large-scaled multi-product production routing problem with outsourcing European Journal of Operational Research 272 914-27
[14] Danna E, Rothberg E and Le Pape C 2005 Exploring relaxation induced neighborhoods to improve MIP solutions Mathematical Programming 102 71-90
[15] D’Andreagiovanni F, Krolikowski J and Pulaj J 2015 A fast hybrid primal heuristic for multiband robust capacitated network design with multiple time periods Applied Soft Computing 26 497-507
[16] Momeni M and Sarmadi M 2016 A genetic algorithm based on relaxation induced neighborhood search in a local branching framework for capacitated multicommodity network design Networks and Spatial Economics 16 447-68
[17] Kamalov F and Thabrah F 2017 A feature selection method based on ranked vector scores of features for classification Annals of Data Science 4 483-502
[18] He F and Zhang L 2018 Prediction model of end-point phosphorus content in BOF steelmaking process based on PCA and BP neural network J. of Process Control 66 51-8
[19] Duc D N and Nananukul N 2018 Supply chain model for renewable energy resource from biomass in Vietnam. Proc. Int. Conf. 2018 5th on Industrial Engineering and Applications (ICIEA): IEEE) pp 520-5
[20] Kostin A, Macowski D H, Pietrobelli J M, Guillén-Gosálbez G, Jiménez L and Ravagnani M A 2018 Optimization-based approach for maximizing profitability of bioethanol supply chain in Brazil Computers & Chemical Engineering 115 121-32