A Mixture of Expert-Based Prediction Approach by Using Genetic Programming and Clustering: A Case Study on Predicting the On-Time Percentages of Local TRA Trains in Taiwan

Chih-Ming Hsu

***Abstract*—The accurate prediction of on-time percentages of trains is an important issue since it can significantly affect the management of trains' operation in many fields, such as making an appropriate timetable, organizing trains' waiting, arranging trains' running tracks, settling the required manpower etc. However, the problem of predicting the on-time percentages of trains is very complex and difficult since the on-time percentages of trains can be influenced by numerous factors. To deal with such a problem, genetic programming (GP) and clustering techniques are used to develop a prediction procedure by means of a mixture of experts in this study. The GP method is utilized to construct prediction models, as well as to identify the feature variables, i.e. critical independent variables. The clustering approach is applied to partition the data into several clusters where data in each cluster are as similar as possible. To illustrate the usefulness and effectiveness of this approach, prediction of the on-time percentages of local trains operated by the Taiwan Railway Administration (TRA) in Taiwan are demonstrated as a case study. The execution results show that the GP can construct an adequate model for predicting the on-time percentages of local trains. Furthermore, a comparison also shows that the technique of mixture of experts can yield a superior GP prediction model by evaluating the performance through MSE, R^2, and MAPE. Hence, we can consider our proposed prediction approach to be a useful and effect procedure for resolving a problem of prediction in the real world.**

***Index Terms*—Mixture of Expert, Genetic Programming, Clustering, On-time Percentage, TRA.**

I. Introduction

Prediction problems exist almost everywhere in our living world, e.g. prediction of stock prices, power consumption, rainfall, etc. For prediction problems involving trains, many methods that apply techniques from different fields have been proposed, and their effectiveness and usefulness have been demonstrated. For example, [1] developed a hybrid model called GAANN that utilizes genetic algorithms (GAs) to optimize the network architecture of artificial neural networks (ANNs) so as to forecast passenger volume in each month on Serbian railways. Hence, the number of neurons in the middle layer of ANNs can be determined by using the selected population in GAs. A time series of the total monthly number of passengers flows gathered from the SORS (the Statistical Office of the Republic of Serbia) was used to assess the predicting performance of the GAANN. A comparison between the GAANN and the traditional SARIMA (Seasonal Autoregressive Integrated Moving Average) model revealed that their proposed approach can yield better forecasting results. [2] developed a model that uses the Holt-Winters model with consideration of the changes in TSF (train service frequency) for the OD (origin to destination) at different times during an operating day to forecast passenger flow in the short term on high-speed rail. The Holt-Winters model can take advantage of characteristics in the time series of passenger flow. In addition, the changes of TSF for the OD at different times in a day are also considered. The final hybrid model is generated through integration based on the minimum absolute value method. To verify the effectiveness, the operational data of the high-speed railway from Beijing to Shanghai railway during 2012 to 2016 were used to demonstrate their proposed model. Furthermore, their method can be further applied to forecast the effects of the TSF with a definite formation. [3] combined temporal forecasting based on a radial basis function neural network (RBF NN) and spatio forecasting based on spatial correlation degree to develop an approach for forecasting the passenger flow status in a high-speed railway transport hub (HRTH). The temporal forecasting based on RBF NN is utilized to forecast passenger flow status in the bottleneck position, as well as combining a spatio forecasting approach based on spatial correlation degree to improve the forecasting precision. The computational experiments on actual passenger flow status of a specific bottleneck position and its correlation points in the Chinese HRTH revealed the effectiveness of their proposed approach in forecasting the passenger flow status with high precision. [4] developed several railway level crossing (LX) accident prediction models that can highlight the main parameters' influences. Based on LX accidents, they utilized the ordinary least-squares (OLS) and nonlinear least-squares (NLS) methods to estimate the respective coefficients for variables in the prediction models. The dedicated accident database provided by SNCF (National Society of French Railway Networks) Réseau was used as a case study to validate the performance of their proposed model, and a comparison process was made through evaluation by statistical means for examining how well their models' estimations can fit the reality. The experimental and compared results prove that their proposed improved accident prediction model can produce statistic-based approbatory quality, while the

Manuscript received July 23, 2019
Chih-Minh Hsu, Department of Business Administration, Minghsin University of Science and Technology, Hsinchu, Republic of China

combination of the proposed model with the negative binomial distribution can yield relatively high accuracy of prediction for the probability of accident occurrence. [5] aimed to develop several primary delay recovery (PDR) predictor models to predict recoverable train delay accurately. They first utilized operation records from the Wuhan-Guangzhou high-speed railway (HSR) to identify the main variables, as well as individual sections' influence, that affect the delay of a train by developing a general framework that can be applied to any HSR line. Random forest regression (RFR), multiple linear regression (MLR), support vector machine (SVM), and artificial neural networks (ANN) are then applied to predict the PDR. The validation results on test datasets showed that the RFR prediction model can outperform the other three alternative models while measuring the prediction accuracy. In addition, the RFR model can achieve a prediction accuracy of more than 80%, while the prediction tolerance is less than one minute based on their evaluation results. [6] applied the neural network and origin-destination (OD) matrix estimation to propose a divide-and-conquer method for forecasting the short-term passenger flow in a high-speed railway system. They first collected the numbers of arriving and departing passengers at each station to form the OD matrices. The neural network was then used to comprehend the short-term forecasting for the arriving/departing passengers' flow. Finally, an OD matrix estimation method was utilized to obtain the OD matrices for a short-term timeframe. Verification through a case study on a high-speed railway with fifteen stations in China showed that the proposed divide-and-conquer method can indeed perform adequately in forecasting the short-term passenger flow on a high-speed railway. [7] applied fuzzy logic relationship recognition techniques to build a fuzzy-temporal-logic-based passenger flow forecast model (FTLPFFM) where the past sequences of passenger flow are considered by using fuzzy logic relationship recognition techniques in the searching process to predict the short-term passenger flow for a high-speed railway. The implementation results using real-world data indicated that the FTLPFFM model can significantly improve the forecast accuracy in terms of measuring with MAE, MAPE, and RMSE compared to the ARIMA (autoregressive integrated moving average model) and KNN (k-nearest neighbor) models. [8] utilized a timed event graph with dynamic arc weights to develop a microscopic model to predict train event times. Through using processed historical track occupation data to reflect all phenomena of railway traffic captured by train describer systems and preprocessing tools, the process times in the model can be dynamically obtained. Next, the characteristics regarding the graph structure of the model allow fast algorithms to be applied to estimate the event times even for large networks. Incorporation of predicted route conflicts on train running times due to braking and reacceleration can further increase the prediction accuracy. Furthermore, the expected prediction error is continuously minimized adaptively by detecting the train runs with process times that successively deviate from their estimates in a certain pattern, as well as the downstream process times. The train describer log files on the busy corridor of Leiden-The Hague-Rotterdam-Dordrecht in the Netherlands were used to test and validate their proposed tool, and adequate experimental results were obtained. However, various errors in logging of event times can still occur even though detailed data with high quality have been used. [9] proposed an approach based on fuzzy rules and time series analysis for predicting online failure in railway transportation systems. They model the relationships among different variables while applying univariate time series analysis to describe the evolution of each variable. Two predicted values can be obtained for a dependent variable, where one variable is produced from the time series model, and the other one is computed from the fuzzy rules by implementing fuzzy inference. A failure in some time period ahead can be declared when the difference between the two values has exceeded a pre-specified threshold. The authors claim that their proposed method differs from the existing methods since it not only considers the evolutionary trend of each variable but also reflects the relationships among different variables. In addition, no prior knowledge of the system model or failure patterns are required. An ATP railway transportation system was used to illustrate their proposed method, and experimental results also proved that the proposed method can predict online system failures effectively. [10] presented a neural network model aiming at predicting the delay of passenger trains of Iranian Railways. They utilized three methods, including normalized real number, binary coding, and binary set encoding, to define inputs. In addition, three different strategies, called quick method, dynamic method, and multiple method, are investigated to find an appropriate architecture for a neural network to a specific task. The data regarding passenger train delays are also divided into three parts, training, validation, and testing sets, to prevent the occurrence of overfitting while modeling a neural network by making cross validation. The execution results by implementing three different data input methods and three different architectures are compared with each other, as well as with well-known prediction methods including the decision tree and multinomial logistic regression. A time-accuracy graph is plotted to make a fair comparison among all models, and the results indicate that their proposed model indeed can yield higher prediction accuracy. [11] exploited big data technologies, learning algorithms, and statistical tools to build a data-driven train delay prediction system (TDPS) for railway networks with large scale. Their intention was to exploit the recent in-memory large-scale data processing technologies for predicting train delays by developing a fast learning algorithm for shallow and deep extreme learning machines. It was demonstrated that their proposed method can perform up to twice as well as the current state-of-the-art methods through a case study and comparison on the train movement data provided by RFI (Rete Ferroviaria Italiana) in the real world. The authors also present methods for tuning the hyperparameters of the learning algorithms efficiently and effectively. In addition, robust models with high performance with respect to the actual train delay prediction system of RFI are derived by deeply exploiting the historical data on train delays. [12] proposed a method of calculating spatial correlation degree between a key area and a correlated surveillance area, and developed an algorithm to forecast passenger flow risk according to the spatial correlation. The effectiveness of their proposed approach was verified through implementing computational experiments on a specific key area in a high-speed railway transport hub, and adequate results were obtained. [13] proposed a booking model by using a framework of case-based reasoning based on reservation data. There are four modules containing distinctive functions for similarity evaluation, instance selection, arrival projection, and parameter search included in his proposed method. The forecasting capability was validated by testing the proposed model on fourteen collected data series and comparing the out-of-sample accuracy based on the four traditional benchmarks. According to the empirical results, his proposed self-learning model could reduce at least 11% of mean square errors (MSE) on average, and the MSE can be significantly reduced through the learning scheme in his proposed approach in comparison with the other naïve versions' prediction performance.

As the above literature review shows, problems regarding the prediction of the on-time percentages of trains has not been investigated previously. However, predicting the on-time percentages of trains is a critical issue for the operation of a railway corporation since the on-time percentages of trains have considerable effects on making an appropriate timetable, organizing the stations for waiting trains, organizing the running tracks of trains, arranging the sufficient manpower etc. However, various factors can influence the on-time percentages of trains, e.g. the total number of passengers, passengers' types, operating days of trains, the initiation and terminal stations or time, the running distances for trains etc., and it is not easy to identify and simultaneously consider all of them completely. In addition, it is also a difficult task to gather the operation data of trains. The prediction for the on-time percentages of trains is therefore considered a very complicated and thorny problem, and has been the subject of far fewer studies. Next, the mixture of experts is a practical technique for resolving a single problem by partitioning the original problem into several sub-problems with smaller sizes, and tackling each sub-problem to create an expert that is dedicated to solving its corresponding specialized sub-problem. Therefore, this study develops a prediction approach by using a technique of a mixture of experts based on genetic programming (GP) and clustering analysis. The remaining sections are organized as follows. The analyzing and modelling methodologies are briefly introduced in Section 2. Then, Section 3 presents our proposed prediction approach. Section 4 gives a real case study on predicting the on-time percentages of local trains operated by the Taiwan Railway Administration in Taiwan. Finally, conclusions are provided in Section 5.

II. Analyzing and Modelling Methodologies

There are two main analyzing and modelling methodologies including the two-step clustering and genetic programming in our proposed prediction approach. This section briefly interprets these two research methods by introducing two-step clustering first.

A. Two-step Cluster Analysis

Cluster analysis is a method which groups a set of data such that the data in the same group, i.e. cluster, are more similar to each other than to other groups' data. Cluster analysis can be achieved by various approaches. TwoStep cluster analysis is a famous one among these methods. The TwoStep cluster analysis procedure utilizes a likelihood distance measure where the variables in the cluster model are assumed to be independent. Furthermore, the distribution for each continuous variable is assumed to be normal, and a multinomial distribution for each categorical variable is assumed. The TwoStep cluster analysis differs from the traditional clustering techniques in that it has several desirable features including: (1) It can create clusters based either on categorical or on continuous variables; (2) It can select the optimal number of clusters automatically; (3) It can analyze large scales of data files efficiently. The procedure of TwoStep cluster analysis is summarized as follows:

Step 1. Construct a cluster features (CF) tree

The cluster features (CF) tree places the first case at the root of the tree in a leaf node where the variable information of that case is contained. Then, each successive case is added to an existing node or it can form a new node according to its similarity to the existing nodes, as well as based on the similarity criterion that uses the distance measure. Therefore, a node can summarize the variable information about the cases clustered in the node. The CF tree hence can give capsule summary information for the data file.

Step 2. Group the leaf nodes

An agglomerative clustering algorithm is applied to group the leaf nodes of the CF tree to produce a range of solutions. Then, Schwarz's Bayesian Criterion (BIC) [14] or the Akaike Information Criterion (AIC) [15] can be used as the clustering criterion to determine the optimal number of clusters.

B. Genetic Programming

Through observing the evolution progress of organisms in the natural world, C. R. Darwin put forward his famous theory of natural selection and evolution. [16] presented the well-known optimization method, called genetic algorithms (GAs), based upon inspiration by Darwin's evolution theory, to solve an optimization problem by imitating the evolutionary procedure of living beings. In GAs, a feasible solution (individual) for an optimization problem is represented by using a series of genes, called a chromosome, that mimics the chromosome of a living thing. All individuals form a population. The excellence of each feasible solution in the population is evaluated by the fitness, assessed according to the fitness function designed based on the objective function in an optimization problem. The mechanism of natural selection and matching is designed to simulate the marriage of individuals to form a matching pool. Then, the paired individuals, called parents, in the matching pool can hopefully produce superior new individuals, called offspring, by using well-designed crossover functions which closely relate to the fitness corresponding to a feasible solution in an optimization problem. In addition, a mutation function is also designed to represent the unusual situation of crossover, i.e. extraordinary genetic changes. Finally, the individuals of offspring are assessed by the fitness function, and the better individuals among the offspring replace the worse (weak) individuals in the previous generation, i.e. their parents' generation, thus forming a new population in the next generation. Later, [17] developed genetic programming (GP), which extends the GAs into a field of computer programs by expressing a feasible solution (program) through a tree-based structure as shown in Figure 1. The tree in Figure 1 represents a computer program; we can decode the tree from left to right, as well as from bottom to top, as follows:

$$3 - \frac{x}{15} + 8 \times \sqrt{y}$$

The tree in GP is made up of elements from two parts including the terminal and function sets. The terminal set defines the elements that are available for each terminal branch of the GP program (chromosome). It can be the independent variables, zero-argument functions, random constants, etc. For example, the 3, *x*, 15, 8, and *y* are the elements in the terminal set. The function set is a set of primitive functions available to each branch of the tree-structure program, such as addition, square root, multiplication, sine and others. The +, -, ×, ÷, and √ are the elements from the function set. The fitness corresponding to the above equation can be evaluated by feeding variables *x* and *y* into Equation (1), as well as referring the objective function that it is intended to optimize. Next, the crossover and mutation operators in GAs can also be transformed to the styles that can fit the tree-based GP as illustrated in Figures 2 and 3, respectively. In Figure 2, the original paired solutions include:

$$3 - \frac{x}{15} + 8 \times \sqrt{y}$$

$$4 + \cos(x) - \frac{\log(y)}{6z}$$

The new paired solutions will be:

$$3 - \cos(x) + 8 \times \sqrt{y}$$

$$4 + \frac{x}{15}\cos(x) - \frac{\log(y)}{6z} \qquad (1)$$

Similarly, the original tree, i.e. $3 - \frac{x}{15} + 8 \times \sqrt{y}$, in Figure 3 mutates into a new program, i.e. $3 - \frac{x}{15} + 8 \times (5 + y)$.

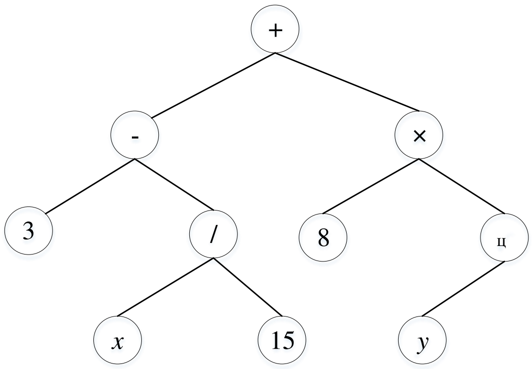

Figure 1. Tree-based genetic programming

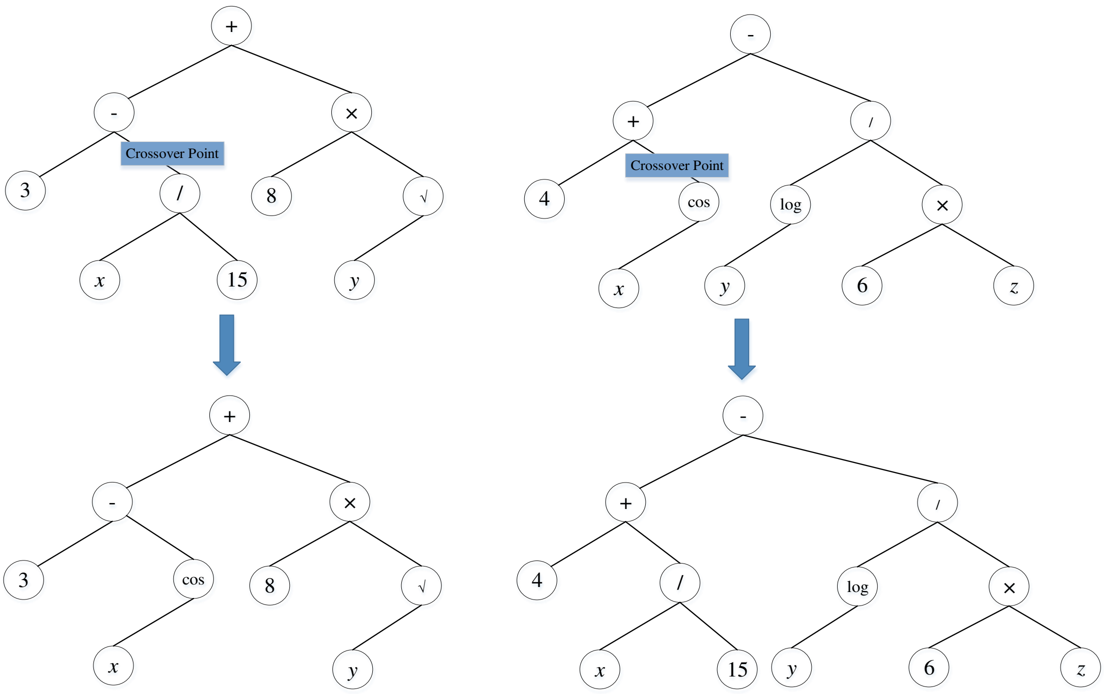

Figure 2. Crossover in the tree-based GP

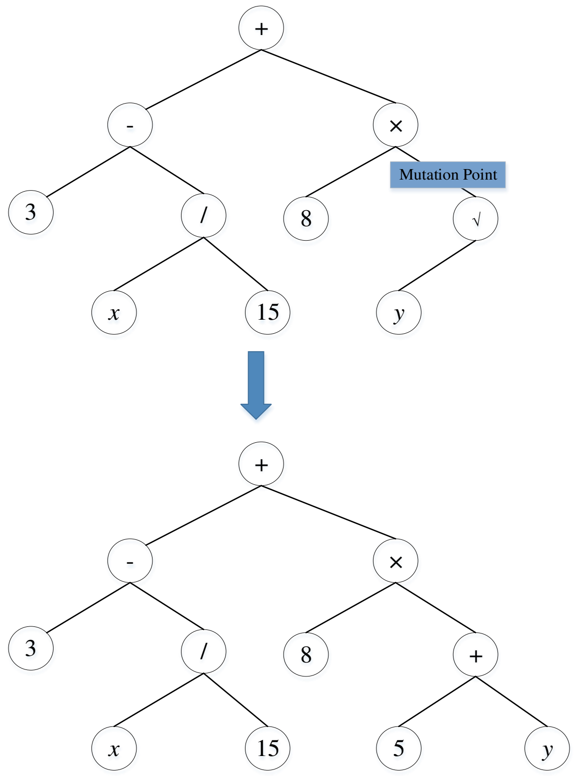

Figure 3. Mutation in the tree-based GP

On the basis of the above definitions, the simple GP can be depicted by Figure 4 and stated by the following procedure [18-20]:

1. Initialization
   The control parameters in GP, e.g. population size, maximum size of programs, crossover rate, mutation rate etc. are first specified. Then, a population of initial solutions (programs or individuals) is generated based on some well-designed mechanism or randomly. In general, the programs should be generated with the condition of a pre-specified maximum size, and the individuals can have different sizes and shapes.
2. Evaluation
   Each program in the population is executed and its fitness (adaptability) in the population is explicitly or implicitly measured in terms of how well it can solve the optimization problem through a pre-defined fitness function. The ways for evaluation can be various, e.g. the amount of error between its output and target, the total cost/time for bringing the system to a desired state, or the classification accuracy. The evaluated result is called the fitness.
3. Create the next generation
   Individuals are selected from the population with a probability that is based on fitness. Then, the genetic operators are applied to the selected individuals (programs), including:
   (1) Reproduction: Copy the selected program to create a new individual.
   (2) Crossover: Randomly recombine chosen parts of paired selected programs (called parents) to form two new individuals (called children) in the offspring.
   (3) Mutation: Randomly mutate a randomly chosen part of the selected program to generate a new offspring individual.
   (4) Architecture-altering: Alter the architecture of the selected program to produce a new offspring individual.

   The individuals in the current population (the now-old generation) are replaced by the individuals in the offspring population based on a certain strategy, e.g. elitist strategy, to create individuals in the new population (the next generation).
4. Check the termination criterion
   The single best program ever encountered during the run (i.e. the best-so-far individual) is designated as the final result when the termination criterion is satisfied. Steps 2 to 4 will run iteratively if the termination criterion cannot be fulfilled.

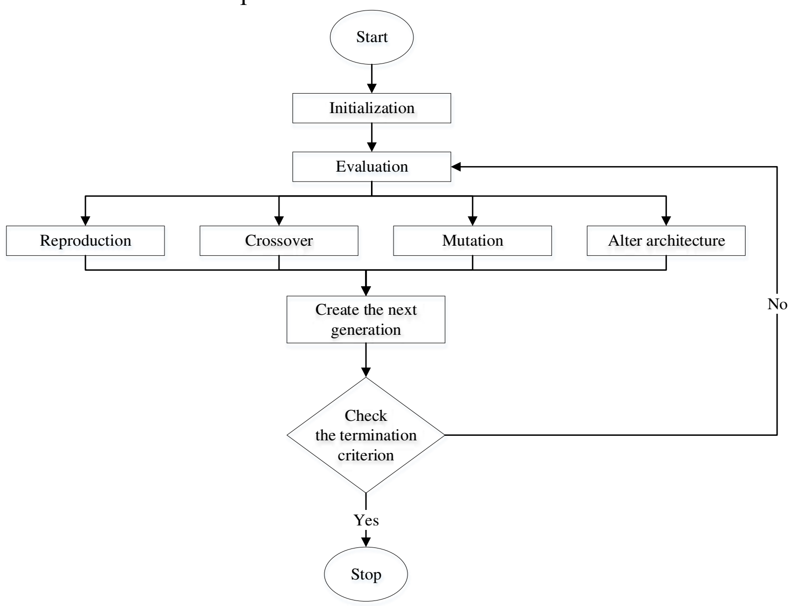

Figure 4. Simple GP flowchart

III. Proposed Approach

This study proposes an approach for predicting the on-time ratios of trains by using the technique of mixture of experts as shown in Figure 5, illustrated as follows:

Step 1: Data preparation

The required data are first collected. To avoid having a variable with a wide range dominate a variable of relatively narrow range, the data in each variable are identically normalized into the range (-1, 1) according to the maximum and minimum values of the corresponding variables. Next, the normalized data are divided into training and test data groups based on a pre-determined ratio, e.g. 3:1. The training data are used to construct a GP prediction model, and the test data are applied to evaluate the generalizability of the well-trained GP model.

Step 2: Construct preliminary GP models

The GP technique is utilized to construct preliminary GP models for predicting the on-time percentages of trains. The prediction performance is measured by the root-mean-square error (RMSE), R squared (R^2), and mean absolute percentage error (MAPE). The GP tool must be implemented several times to obtain multiple prediction models.

Step 3: Select features form preliminary GP models

This step selects features, i.e. important variables that have great impacts on the on-time ratios of trains, according to the preliminary GP models built in the previous step. Each preliminary GP model constructed in Step 2 is analyzed to identify which variables appeared very frequently in its building progress, called appearing variables in this study. Therefore, the featured variables that significantly influence the on-time ratios of trains can be identified through synthesizing the information of appearing variables of each GP model built in Step 2.

Step 4: Divide data into sub-problems

The TwoStep cluster analysis is used where the featured variables identified in Step 3 act as the bases for clustering. The optimal number of clusters is determined according to Schwarz's Bayesian Criterion (BIC) (Burnham and Anderson, 2002) or the Akaike Information Criterion (AIC) (Akaike, 1969) criterion. Hence, the original data prepared in Step 1 can be segmented into several sub groups. The data of each sub-group form a new prediction problem.

Step 5: Construct GP models for each sub-problem

As in Step 1, the data in each sub-group clustered in Step 4 are partitioned into training and test data groups according to a pre-specified ratio. Then, the GP method is applied to the training data to establish several GP prediction models, and the model with the optimal prediction performance, measured by RMSE, R^2, and MAPE, is designated to be the final GP prediction model. In addition, the generalizability of the final GP model to unknown data is evaluated by predicting the on-time ratios of trains in the corresponding test data group for each sub-problem.

Step 6: Mix GP models of each sub-problem into a final model

In Step 5, the final GP model of each sub-problem can be viewed as an expert for predicting on-time percentages of trains for each corresponding sub-problem. Hence, each final GP model selected in each sub-problem is then combined to make an integrated GP prediction model that can predict on-time percentages of trains for all sub-problems, i.e. all of the original data.

Step 7: Evaluate prediction performance

The original GP model built in Step 2 and the integrated GP model combined in Step 6 are compared to contrast their prediction performance, as well as to confirm the superiority of the integrated GP. The criteria for evaluating prediction performance can include MSE, R^2, MAPE etc.

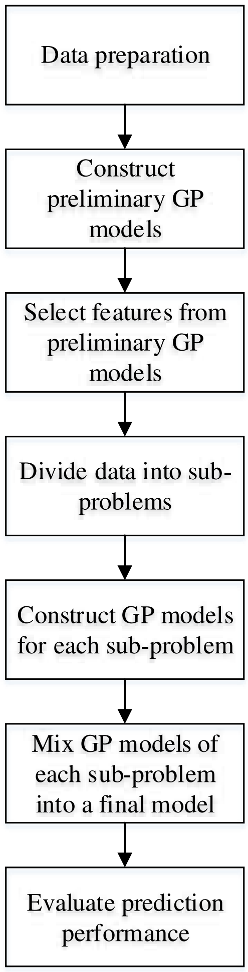

Figure 5. Proposed prediction approach

IV. Case Study

In this section, a case study on prediction of on-time ratios of trains is presented to demonstrate the usage of our proposed approach.

A. Introduction to TRA

Taiwan Railway Administration (TRA) is a traditional railway transportation company owned and operated by the government of the Republic of China (ROC). The first railway was constructed in 1891 with 28.6km of track. Now, the total length of railways operated by TRA has reached

1065km, and contains twelve lines as shown in Table 1 and Figure 6. Since the population concentrates mainly in the western areas, especially in the northwest areas, the traffic of railways between Zhunan and Keelung is considered the busiest. In addition, there are several types of passenger trains, e.g. Tzu-Chiang Limited Express, Puyuma Express, Taroko Express, Chu-Kuang Express, Fu-Hsing Semi Express, Fast Local, and Local Trains operated simultaneously on the same tracks. Among these trains, the Local Trains must often wait for an Express or Fast Local Train to pass, thus the on-time percentages of Local Trains are influenced by many factors and are not easily predicted. Therefore, this study focuses on predicting the on-time percentages of Local Trains which operate covering the region between Zhunan and Keelung stations.

Table 1. TRA operation lines

No.	Lines
1	Chengzhui Line
2	Eastern Trunk Line
3	Jiji Line
4	Liujia Line
5	Neiwan Line
6	Pingxi Line
7	Shenao Line
8	Shalun Line
9	South Link Line
10	Suaoxin-Suao Line
11	Western Trunk Line
12	Western Trunk Line (Coast Line)

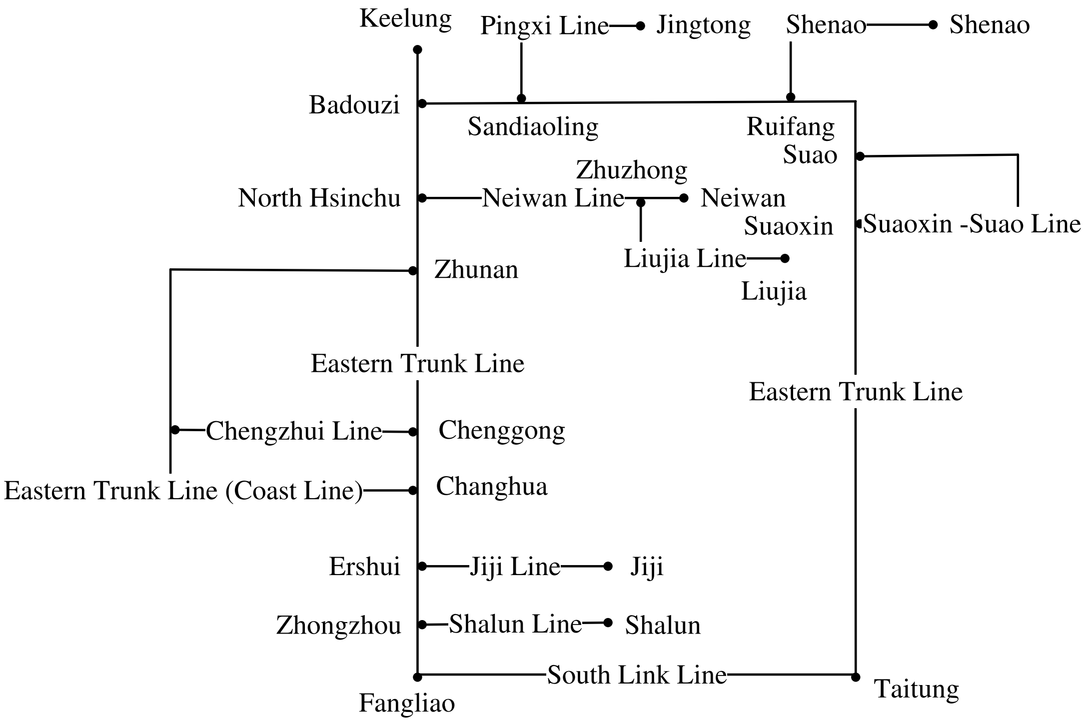

Figure 6. Map of TRA operation lines

B. Data Preparation

First, several variables that are thought to be possible factors, i.e. independent variables, that might affect the on-time percentages of Local Trains are first determined, as summarized in Table 2 (also refer to Figure 6). The response, i.e. dependent variable, is the on-time percentage of each Local Train. The on-time percentage of a Local Train is defined as the total number of trains that can arrive at the terminal station on time divided by the total number of trains that arrive at the terminal station during an operating period, usually one month. Furthermore, a train is deemed to be on time if the train can arrive at its terminal station within five minutes of its scheduled time. For each Local Train, the thirty-three prediction variables, i.e. independent variables, as shown in Table 2 along with its corresponding on-time percentage, i.e. dependent variable, are arranged into a row. Based on the operation data collected by TRA on March 2019, this study gathered the data including 278 rows. Among these independent variables, the ranges of some predictors are large relative to certain other ones. For example, the minimum and maximum values for the “Initial time” (independent variable #5) are 100 and 1400, respectively. However, the “Train-kilometer” (independent variable #6) only ranges from 2.1 to 130.9. In order to avoid the effects of the large-range variables dominating the effects of variables with relatively small ranges on the dependent variable, all variables (including the independent and dependent variables) in the collected data with 278 rows are first normalized linearly to a range from -1 to 1 based on their corresponding maximum and minimum values. The normalized data are then randomly partitioned into the training and test data sets with 209 and 69 rows, respectively, based on a ratio of 3:1.

Table 2. Possible factors that might affect the on-time percentages of Local Trains

No.	Factors (Independent variables)	Explanation
1	Northbound/Southbound	The train is northbound/southbound.
2	Operation days	Train runs daily/on Sundays/daily except Sundays/daily except Saturdays.
3	Operation lines	Train runs via Eastern Trunk line/ Eastern Trunk Line (Coast Line)/Cross-Line (Mixed Lines).
4	Initial kilometer marker	Kilometer marker ("kmarker") of station from which train initiates. Zhunan station kmarker is set at 0. Kmarker of each station is calculated based on distance from Zhunan station.
5	Initial time	Time when train sets out.
6	Train-kilometers	Distance of train's trip.
7	Traveling time	Duration of train's trip.
8	Train-kilometers before train enters region	Northern point at which Western trunk and Western trunk line (coast line) merge is Zhunang station. Eastern trunk line starts at Badouzi station. Trains not setting out from points between Zhunang and Keelung stations have run for a distance before entering region between Zhunang and Badouzi stations. This distance is defined as "Train-kilometers before train enters region".
9	Travel time before train enters region	Similar to explanation in previous factor. Trains not initiating between Zhunang and Keelung stations have run for a duration before entering region between Zhunang and Badouzi stations. This duration is defined as "Travel time before train enters region".
10	Number of waiting times before train enters the region	Similar to situation explained in factor 8. Trains not initiating between Zhunang and Keelung stations must wait for Express/Fast Local Train several times before it enters region between Zhunang and Badouzi stations. This *number of* waiting times is defined as "Number of waiting times before train enters region".
11	Waiting time before train initiates	Similar to situation explained in factor 8. A train not initiating between Zhunang and Keelung stations must wait for Express/Fast Local Train for a duration before it enters region between Zhunang and Badouzi stations. This waiting *duration* is defined as "Waiting time before train initiates".
12	Train-kilometers after train leaves region	Similar to situation explained in factor 8. A train not terminating between Zhunang and Keelung stations must run for a distance after leaving region between Zhunang and Badouzi stations. This distance is defined as "Train-kilometers after train leaves region".
13	Traveling time after train leaves region	Similar to situation explained in factor 8. A train not terminating between Zhunang and Keelung stations must run for a duration after it leaves region between Zhunang and Badouzi stations. This duration is defined as "Traveling time after train leaves region".
14	Number of waiting times after train leaves region	Similar to situation explained in factor 8. A train that does not terminate between Zhunang and Keelung stations must wait for Express/Fast Local Train several times after it leaves region between Zhunang and Badouzi stations. This number of waiting times is defined as "Number of waiting times after train leaves region".
15	Waiting time after train leaves region	Similar to situation explained in factor 8. A train that does not initiate between Zhunang and Keelung stations must wait for Express/Fast Local Train for some duration after it leaves region between Zhunang and Badouzi stations. This waiting duration is defined as "Waiting time after train leaves region".
16	Number of waiting times during operating (Tzu-Chiang Limited Express)	The total number of delays due to waiting for Tzu-Chiang Limited Express to pass while running between Zhunang and Keelung stations.
17	Number of passing trains during operating (Tzu-Chiang Limited Express)	Total number of passing Tzu-Chiang Limited Express trains while a train runs between Zhunang and Keelung stations.
18	Operating distance of awaited Express (Tzu-Chiang Limited Express)	Total distance which passing Tzu-Chiang Limited Express trains have run before Tzu-Chiang Limited Express passes a train that runs between Zhunang and Keelung stations.
19	Waiting time of a train (Tzu-Chiang Limited Express)	The total time a train must wait for Tzu-Chiang Limited Express to pass while running between Zhunang and Keelung stations.
20	Number of waiting times during operating (Puyuma/Taroko Express)	The total number of delays due to waiting for Puyuma/Taroko Express to pass while running between Zhunang and Keelung stations.
21	Number of passing trains during operating (Puyuma/Taroko Express)	Total number of passing Puyuma/Taroko Express trains while a train runs between Zhunang and Keelung stations.
22	Operating distance of awaited Express (Puyuma/Taroko Express)	Total distance passing Puyuma/Taroko Express trains have run before Puyuma/Taroko Express passes a train that runs between Zhunang and Keelung stations.
23	Waiting time of a train (Puyuma/Taroko Express)	Total time a train must wait for Puyuma/Taroko Express to pass while running between Zhunang and Keelung stations.
24	Number of waiting times during operating (Chu-Kuang Express)	Total number of delays while a train must wait for Chu-Kuang Express to pass while running between Zhunang and Keelung stations.
25	Number of passing trains during operating (Chu-Kuang Express)	Total number of passing Chu-Kuang Express trains while a train runs between Zhunang and Keelung stations.
26	Operating distance of awaited Express (Chu-Kuang Express)	Total distance the passing Chu-Kuang Express trains have run before Chu-Kuang Express pass a train that runs between Zhunang and Keelung stations.
27	Waiting time of a train (Chu-Kuang Express)	Total waiting time for a train while waiting for Chu-Kuang Express to pass while running between Zhunang and Keelung stations.
28	Number of waiting times during operating (Fast Local Train)	Total number of delays while a train must wait for Fast Local Train to pass while running between

		Zhunang and Keelung stations.
29	Number of passing trains during operating (Fast Local Train)	Total number of passing Fast Local Trains while a train runs between Zhunang and Keelung stations.
30	Operating distance of awaited Express (Fast Local Train)	Total distance which passing Fast Local Trains have run before Fast Local Trains pass a train that runs between Zhunang and Keelung stations.
31	Waiting time of a train (Fast Local Train)	Total waiting time for a train waiting for Fast Local Train to pass while running between Zhunang and Keelung stations.
32	Number of waiting times for awaiting one train	Total number of delays while a train must wait for any type of train while running between Zhunang and Keelung stations.
33	Number of waiting times for successively awaiting two trains	The total number of delays when a train must successively wait for two of any type of train while running between Zhunang and Keelung stations.

C. Build Preliminary GP Prediction Models

The GP technique is then applied to the prepared training and test data sets to build preliminary GP prediction models for the on-time percentages of trains. Here, the Discipulus GP software is used with its default parameter settings. The important parameters, including the population size, crossover rate, and mutation rate, are set as 500, 0.5, and 0.95, respectively. The GP software is repeatedly run ten times to select the optimal prediction model, and Table 3 summarizes the execution results. The fifth execution, denoted by a star (*), is selected as the best preliminary GP model since it can yield relatively low training RMSE and MAPE, as well as high training R^2 among ten runs. In addition, the coefficient of variance (CV) for the RMSE, R^2, and MAPE are 0.119279, 0.03201, and 0.14087, respectively, for the training data set. These values can be considered low enough. Therefore, the preliminary prediction models built through GP can be thought of as adequately robust.

Table 3. Execution results for building preliminary GP models

Execution No.	Training			Test		
	RMSE	R^2	MAPE	RMSE	R^2	MAPE
1	0.049533	0.71775	0.14033	0.047474	0.71523	0.11198
2	0.040292	0.70159	0.11509	0.056866	0.59651	0.12092
3	0.042408	0.69721	0.12253	0.057540	0.61642	0.13027
4	0.045794	0.69648	0.13509	0.055894	0.62280	0.11493
5*	0.033239	0.76070	0.10278	0.026404	0.81803	0.08709
6	0.044826	0.71107	0.13180	0.059220	0.62012	0.13106
7	0.036971	0.74628	0.14069	0.034618	0.75413	0.09325
8	0.037352	0.73539	0.13647	0.031590	0.79798	0.09385
9	0.046159	0.69462	0.13734	0.059839	0.60852	0.12654
10	0.041119	0.72748	0.12116	0.059599	0.58558	0.13238
Mean	0.041769	0.71885	0.12833	0.048904	0.67353	0.11423
Standard Deviation	0.004982	0.02301	0.01254	0.013078	0.08893	0.01720
CV	0.119279	0.03201	0.14087	0.267425	0.13204	0.13143

D. Feature Selection Based on Preliminary GP Models

The frequency of appearance for each independent variable while evolving in the building processes of the best GP model, i.e. the fifth model shown in Table 3, is summarized in Table 4. Based on Table 5, the independent variables #2 and #5 always appear, i.e. a 100% frequency, in the evolving process of the best preliminary GP model. Therefore, these two independent variables are considered to be the most critical factors that can influence the dependent variable, i.e. on-time percentage of a train, and are thus selected as the featured variables in the TwoStep cluster analysis shown in the following Section.

Table 4. Frequencies of appearance of each independent variable while evolving the best GP model

Independent variable no.	1	2	3	4	5	6	7	8	9	10	11
Appearance frequency	50%	100%	50%	67%	100%	33%	17%	0%	17%	50%	50%
Independent variable no.	12	13	14	15	16	17	18	19	20	21	22
Appearance frequency	40%	30%	17%	50%	23%	10%	50%	83%	33%	50%	67%
Independent variable no.	23	24	25	26	27	28	29	30	31	32	33
Appearance frequency	23%	3%	60%	17%	37%	0%	3%	33%	33%	20%	13%

E. TwoStep Cluster Analysis for Dividing Data

The two independent variables identified in Section 4.D are fed into the TwoStep cluster analysis for further clustering the original data prepared in Section 4.B to yield data groups with enough diversity. SPSS software is used to implement the TwoStep cluster analysis where the distance between two items are evaluated by the likelihood measure, the BIC (Schwarz's Bayesian Criterion) clustering criterion is applied, and the optimal number of clusters is determined automatically by the SPSS process. Hence, the original data prepared in Step 1 with 278 rows are divided into four sub-groups. In addition, the data in each sub-group cluster are further partitioned into training and test data groups based on a pre-determined 3:1 ratio as shown in Table 5. Table 5 also depicts the clustering centers of independent variables #2 and #5 for each group's data. From Table 5, the clustering centers of independent variables #2 and #5 in the first sub-group are 1.39 and 760.38, respectively. The second and fifth independent variables (factors) are "Operation days" and "Initial time" as shown in Table 2. In addition, the "Operation days" are coded as 1, 2, 3 and 4 for the trains running daily, on Sundays, daily except Sundays, and daily except Saturdays. Meanwhile, the values of "Initial time" are coded according to the time when the train sets out, e.g. coding 8:30

and 15:20 as 510 (minutes) and 920 (minutes), respectively. Therefore, the first sub-group's data mainly represent the Local Trains that run daily or on Sundays, and depart around 12:30. Similarly, the second sub-group's data primarily represent the Local Trains that run on Sundays and initiate about 15:30. The major Local Trains that run on Sundays and start from their initiation station at about noon are partitioned into the third sub-group. Finally, the last sub-group's data mainly stand for the Local Trains with running days of Sundays, as well as initiating time of about 13:30.

Table 5. Four sub-groups' data obtained by the TwoStep cluster analysis

Sub groups	Quantity of data	Quantity of training data	Quantity of test data	Clustering center	
				Independent variable # 2	Independent variable # 5
1	76	57	19	1.39	760.38
2	88	66	22	1.70	925.12
3	52	39	13	1.83	732.77
4	62	47	15	1.68	806.85

F. Build GP models for sub-problems

The GP method implemented by the Discipulus software is utilized again for the training and test data of each sub-group shown in Table 5. For each sub-group's data, the GP software is run ten times, and Table 6 summaries the implementation results. Based on the RMSE, R^2, and MAPE, the sixth, ninth, seventh, and second GP models in Table 6 are selected as the best prediction models for the four sub-groups' data and denoted by a star (*). From Table 6, it can be seen that the GP tool can build a superior model for predicting the on-time percentages of trains by the training data in the fourth sub-group. In other words, the on-time percentages of the Local Trains that mainly run on Sundays and set out at about 13:30 are relatively easy to predict. The first major reason we consider is that the time needed for getting on and off the trains can be significantly reduced since there are fewer working commuters at noon each day, especially on Sundays. The other reason is that most passengers on Sundays are travelers who start their travelling early in the morning and return home in the evenings or even very late. Hence, this situation further reduces the time for getting on and off the trains, especially around 13:30. Due to the above reasons, the on-time percentages of trains are higher, and thus are relatively easily predicted. Notably, the prediction performance via the tests MSE, R^2, and MAPE are not always the best among all of the best GP models selected for each sub-group's data. However, the test R^2 still can attain a value of 0.89, which is considered high enough in the field of social science. On the other hand, the training MSE, R^2, and MAPE in the second sub-group are worse among the best GP models selected for four sub-groups' data. That is to say that the on-time percentages of Local Trains that run on Sundays and mainly depart about 15:30 are relatively difficult to predict. The main possible cause is that the time 15:30 may be considered a rush time on Sundays since many travelers start their journey to the big cities for fun in the evening. In addition, there are more accidents at the level rail crossings since the road traffic is comparatively busy, thus delaying the running of trains and decreasing the on-time percentages of trains. Therefore, the on-time percentages of trains can vary greatly and are hard to predict. Besides, the MSE, R^2, and MAPE for the test data in the second sub-group are worse among all of the best GP models chosen for each sub-group's data. However, the R^2 for the training and test data can also reach 0.91 and 0.89, respectively, which can be regarded as high. The training and test MAPEs are all less than 15%, considered a difficult level that can be attained in dealing with the difficult prediction problems for on-time percentages of trains. In general, the GP technique can yield prediction models with sufficient quality since the lowest training and test R^2s for all of the best GP models determined in each sub-group can be up to 0.91 and 0.89, respectively, and the largest training and test MAPEs are less than 0.051 and 0.15, individually. Furthermore, the superior training and test R^2s can even achieve levels of 0.98 and 0.97, respectively. Based on the above information, the GP tools can produce a model with sufficient prediction performance, since GP can concentrate on constructing a prediction model for each sub-group's data where the data represent a particular, i.e. similar, situation of running and initiation for Local Trains.

G. Mixing Expert Models and Evaluating Prediction Performance

Each of the selected GP models can be thought of as an expert for predicting on-time percentages of trains for its corresponding sub-group's data. Hence, these selected models are merged to build an integrated prediction model to predict on-time percentages of trains, where which GP model should be applied for prediction is in accordance with which group a train's information is clustered in. Then, the criterion for evaluating prediction performance including MASE, R^2, and MAPE is compared for the selected best preliminary GP model built in Section 4.C, denoted by GP_{All}, and the integrated one combined in Section 4.G, denoted by GP_{Int}. The comparison is provided in Table 7, in which we can see that the GP_{all} prediction model by mixing an expert in its professional field can provide less MSEs and MAPEs, and can yield higher R^2s for both the training and test data compared to the GP_{All} model. In other words, the process for dividing the data by the critical factors, i.e. important feature variables for the dependent variable, into appropriate groups such that each group's data can be similar as much as possible is an effective method for improving the prediction performance of built GP models.

Table 6. GP Implementation Results for Each Sub-group in Table 5

Sub	Run No.	Training	Test

group		MSE	R^2	MAPE	MSE	R^2	MAPE
1	1	0.003878	0.96666	0.03296	0.005185	0.90286	0.03582
	2	0.003746	0.96099	0.03253	0.005211	0.90142	0.03450
	3	0.004360	0.96633	0.03477	0.004385	0.91813	0.02577
	4	0.003718	0.96580	0.03199	0.004761	0.89982	0.03009
	5	0.003832	0.96298	0.02790	0.006025	0.88531	0.03625
	6*	0.002469	0.97867	0.02422	0.002547	0.96548	0.02284
	7	0.002990	0.97303	0.02711	0.007858	0.84500	0.04110
	8	0.002827	0.97185	0.02744	0.004092	0.93712	0.03122
	9	0.003543	0.96940	0.03173	0.007134	0.87358	0.03508
	10	0.004142	0.96866	0.03512	0.007286	0.87669	0.04222
	Mean	0.003551	0.96844	0.03058	0.005448	0.90054	0.03349
	Standard deviation	0.000602	0.00514	0.00364	0.001644	0.03413	0.00615
	CV	0.169647	0.00531	0.11966	0.301858	0.03790	0.18366
2	1	0.012843	0.90383	0.06209	0.058458	0.69453	0.15926
	2	0.014286	0.89381	0.06557	0.038088	0.82979	0.12412
	3	0.013526	0.88571	0.06027	0.043950	0.81992	0.13628
	4	0.013570	0.91142	0.06260	0.048038	0.80901	0.14837
	5	0.011713	0.89068	0.05353	0.045687	0.84975	0.14093
	6	0.013576	0.87850	0.06038	0.043168	0.86403	0.14026
	7	0.013256	0.90287	0.06194	0.026906	0.86186	0.11020
	8	0.011845	0.91103	0.05677	0.040638	0.74573	0.12779
	9*	0.009452	0.91965	0.05076	0.046447	0.89437	0.14449
	10	0.011334	0.89636	0.05632	0.046236	0.77562	0.14287
	Mean	0.012540	0.89939	0.05902	0.043762	0.81446	0.13746
	Standard deviation	0.001450	0.01275	0.00457	0.008012	0.06078	0.01378
	CV	0.115611	0.01417	0.07739	0.183088	0.07462	0.10024
3	1	0.005389	0.97013	0.03990	0.016691	0.96184	0.07177
	2	0.006853	0.96211	0.05026	0.019167	0.95092	0.07839
	3	0.004169	0.97596	0.03726	0.019048	0.96048	0.07374
	4	0.005335	0.97112	0.04741	0.020214	0.95013	0.06846
	5	0.007032	0.95792	0.05405	0.017480	0.96623	0.07762
	6	0.007537	0.95537	0.04837	0.026827	0.93644	0.08446
	7*	0.003667	0.97795	0.03509	0.013700	0.97373	0.04322
	8	0.004750	0.97135	0.03592	0.019509	0.94273	0.09212
	9	0.006794	0.96258	0.04597	0.028211	0.91973	0.09926
	10	0.007558	0.95729	0.04541	0.023546	0.93591	0.08203
	Mean	0.005908	0.96618	0.04396	0.020439	0.94981	0.07711
	Standard deviation	0.001426	0.00813	0.00653	0.004516	0.01643	0.01513
	CV	0.241359	0.00841	0.14851	0.220948	0.01729	0.19617
4	1	0.010539	0.94065	0.05874	0.024060	0.84735	0.08848
	2*	0.003375	0.98444	0.03584	0.011314	0.89106	0.05107
	3	0.006297	0.96767	0.04703	0.016395	0.84859	0.06743
	4	0.010592	0.94088	0.06043	0.023018	0.79539	0.08682
	5	0.009547	0.94642	0.05068	0.017296	0.82439	0.06940
	6	0.009285	0.95911	0.06679	0.014762	0.87253	0.06748
	7	0.006989	0.96676	0.05673	0.014304	0.85150	0.05809
	8	0.005367	0.96852	0.04227	0.017012	0.82408	0.07160
	9	0.005757	0.97611	0.05328	0.019887	0.83396	0.07857
	10	0.009723	0.96411	0.06983	0.011358	0.91817	0.05975
	Mean	0.007747	0.96147	0.05416	0.016940	0.85070	0.06987
	Standard deviation	0.002513	0.01474	0.01061	0.004360	0.03561	0.01211
	CV	0.324319	0.01533	0.19585	0.25746	0.04186	0.17326

Table 7. Comparison of Predicting On-time Percentages of Trains

GP model	Training			Test		
	MSE	R^2	MAPE	MSE	R^2	MAPE
GP_{All}	0.033239	0.76070	0.10278	0.026404	0.81803	0.08709

GP_{Int}	0.005102	0.96273	0.03724	0.020551	0.89427	0.07267

V. Conclusions

The on-time percentages of trains could significantly affect the operation of trains, as they are valuable information for arranging appropriate timetables, determining waiting of trains, deciding the tracks for running trains, and setting up adequate manpower etc. The prediction of the on-time percentages of trains, however, is a very complicated and difficult problem due to various factors that can influence the on-time percentages of trains and that are not easy to completely identify and consider. In addition, segmenting an original problem into several sub-problems, and creating an expert that has expertise in its specialized area by solving each sub-problem, called the mixture of experts, has been popular. Therefore, this study proposes a prediction procedure employing a mixture of experts by using genetic programming (GP) and clustering analysis. The usefulness and effectiveness of the proposed approach are verified through a case study predicting the on-time percentages of Local Trains operated by the Taiwan Railway Administration (TRA) in Taiwan. The implementation results show that GP can adequately construct an original prediction model for the whole dataset. Appropriate groups of data are then found by clustering with characteristics, i.e. features, variables that are identified through exploring the well-constructed GP model. An integrated prediction model can be obtained by mixing experts dedicated to predicting data in each cluster. A comparison shows that the integrated GP prediction model is much better than the original one based on the performance evaluation through MSE, R^2, and MAPE. In other words, the process for building a distinct GP prediction model for each cluster's data, having similar characteristics, can indeed improve the prediction accuracy process. Therefore, our proposed approach can be considered a useful and effective tool for resolving a prediction problem in the real world.

Acknowledgment

The author thanks Jackson Liu in the Secretariat of Taiwan Railway Administration (TRA) in Taiwan, R.O.C. for fully supporting this study, as well as Minghsin University of Science and Technology, Taiwan, R.O.C., for partially supporting this study under Contract No. MUST-108BA-1.

References

[1] N. Glisovic. (2016). A hybrid model for forecasting the volume of passenger flows on Serbian railways. *Oper. Res. 16(2).* pp. 271—285.
[2] Q. Lai, J. Liu, Y. Luo, M. Minshu. (2017). A hybrid short-term forecasting model of passenger flow on high-speed rail considering the impact of train service frequency. *Math. Probl. Eng.* Article ID 1828102.
[3] Z. Xie, L. Jia, Y. Qin. (2013). A hybrid temporal-spatio forecasting approach for passenger flow status in Chinese high-speed railway transport hub. *Discret. Dynam. Nature. Soc.* Article ID 239039.
[4] C. Liang, M. Ghazel, O. Cazier, E. El-Koursi. (2018). Developing accident prediction model for railway level crossings. *Safety. Sci. 101.* pp. 48—59.
[5] C. Z. Jiang, P. Huang, J. Lessan, L. P. Fu, C. Wen. (2019). Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression. *Can. J. Civil. Eng. 46(5).* pp. 353—363.
[6] M. Q. Xie, X. M. Li, W. L. Zhou, Y. B. Fu. (2014). Forecasting the short-term passenger flow on high-speed railway with neural networks. *Comput. Intel. Neurosc.* Article ID 375487.
[7] F. Dou, L. M. Jia, L. Wang, J. Xu, Y. K. Huang. (2014). Fuzzy temporal logic based railway passenger flow forecast model. *Comput. Intel. Neurosc.* Article ID 950371.
[8] P. Kecman, R. M. P. Goverde. (2015). Online data-driven adaptive prediction of train event times. *IEEE. T. Intell. Transp. 16(1).* pp. 465—474.
[9] Z. Ding, Y. Zhou, G. G. Pu, M. C. Zhou. (2018). Online failure prediction for railway transportation systems based on fuzzy rules and data analysis. *IEEE. T. Reliab. 67(3).* pp. 1143—1158.
[10] M. Yaghini, M. M. Khoshraftar, M. Seyedabadi. (2013). Railway passenger train delay prediction via neural network model. *J. Adv. Transport. 47(3).* pp. 355—368.
[11] L. Oneto, E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, D. Anguita. (2018). Train delay prediction systems: a big data analytics perspective. *Big. Data. Res. 11.* pp. 54—64.
[12] Z. Y. Xie, Y. Qin. (2016). A passenger flow risk forecasting algorithm for high-speed railway transport hub based on surveillance sensor networks. *J. Sensors.* Article ID 564790.
[13] T. T. Tsai. (2014). A self-learning advanced booking model for railway arrival forecasting. *Transport. Res. C. 39.* pp. 80—39.
[14] K. P. Burnham, D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag, 2002.
[15] H. Akaike. (1969). Fitting autoregressive models for prediction. *Ann. I. Stat. Math. 21.* pp. 243—247.
[16] J. H. Holland, *Adaptation in Nature and Artificial Systems*. Ann Arbor, MI: The University of Michigan Press, 1975.
[17] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, Mass: MIT Press, 1992.
[18] J. R. Koza, M. A. Keane, M. J. Streeter, W. Mydlowec, J. Yu, G. Lanza, *Genetic Programming: Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. New York: Springer-Verlag, 2005.
[19] I. Ciglarič, A. Kidrič. (2006). Computer-aided derivation of the optimal mathematical models to study gear-pair dynamic by using genetic programming. *Struct. Multidiscip. O. 32(2).* pp. 153—160.
[20] J. R. Koza, M. J. Streeter, M. A. Keane. (2008). Routine high-return human-competitive automated problem-solving by means of genetic programming. *Inform. Sciences. 178(23).* pp. 4434—4452.

Chih-Ming Hsu Chih-Ming Hsu is currently a Professor in the Department of Business Administration at Minghsin University of Science and Technology, Taiwan. He holds a PhD in Industrial Engineering and Management from National Chiao Tung University, Taiwan. His present research interests include quality engineering, optimization methods in industrial applications and data mining applications in CRM.