A hybrid genetic algorithm and linear regression for prediction of NOx emission in power generation plant

Muhammad Afif Bunyamin, Keem Siah Yap, Nur Liyana Afiqah Abdul Aziz, Sheih Kiong Tiong, Shen Yuong Wong, Md Fauzan Kamal

1 Department of Electronics and Communication Engineering, College of Engineering, Universiti Tenaga Nasional
2 Environment Unit, TNB Research Sdn. Bhd.

Email: yapkeemuniten.edu.my

Abstract. This paper presents a new approach of gas emission estimation in power generation plant using a hybrid Genetic Algorithm (GA) and Linear Regression (LR) (denoted as GA-LR). The LR is one of the approaches that model the relationship between an output dependant variable, y, with one or more explanatory variables or inputs which denoted as x. It is able to estimate unknown model parameters from inputs data. On the other hand, GA is used to search for the optimal solution until specific criteria is met causing termination. These results include providing good solutions as compared to one optimal solution for complex problems. Thus, GA is widely used as feature selection. By combining the LR and GA (GA-LR), this new technique is able to select the most important input features as well as giving more accurate prediction by minimizing the prediction errors. This new technique is able to produce more consistent of gas emission estimation, which may help in reducing population to the environment. In this paper, the study’s interest is focused on nitrous oxides (NOx) prediction. The results of the experiment are encouraging.

1. Introduction

The emissions of the acidifying substances are increasing from time to time [1]. A lot of efforts have been taken to measure and to reduce these acidifying gases. Electricity productions sectors produce tons of these harmful gases due to the gas and fossil fuel combustion for the electricity generations. Carbon oxides, sulphur oxides, and nitrous oxides are the main source of the acidifying emissions produced [2]. Previously at power plant, a conventional method is used. The gases emissions are measured by using sensors. However, the costs and the inefficiency occur during the maintenance duration bring the needs of artificial intelligence solution [3].

When other methods may have faced difficulties in solving the problems as number of input features (parameter of power generation plant) maybe huge, Genetic Algorithm (GA), with its independency characteristic to the specific knowledge of the problem, is used to ease the solutions [4] [5]. The main objective of GA is to get the least number of selected features with the minimum percentage error. Once GA selected the feature, the Linear Regression (LR) will be used as prediction mode. This paper is organized as follows: Section 2 describes both algorithms GA and LR. The results of both LR and GA-LR are presents and discuss in Section 3. Section 4 presents the conclusion and suggestion of further works.
2. The hybrid GA-LR model

In the LR model, the dependant variable \( y_i \) is assumed linear to the regression vector \( X_i \) with the assumption of \( n \) data points \( \{y_i, X_i\} \) for. One popular LR equation is in the form of \( y_i = X_i \beta_i \), where \( X_i \) is the input variables which in independent while the dependant variable \( y_i \) also known as response variable which the best set for the data points is produced [6]. The \( \beta_i \) stands for regression coefficient represent the inference of the function [7]. LR is one of the methods that estimate data by using linear predictor functions and unknown model parameters [8]. The solutions of LR can be found by moore-penrose pseudo inverse matrix method [9].

Genetic Algorithms (GA) have been intensively studied during the past few decades and many applications have benefited from the use of GA [10]. GA is adaptive and robust computational methods that can be used to solve search and optimization problems. The dynamics of GA based on the mechanics of natural selection and genetic processes of living organisms. Any GA starts with a population of randomly generated solutions (chromosomes) and advances toward better solutions by applying genetic operators, modeled on genetic processes occurring in nature. From one generation to another, populations evolve according to the principles of natural selection and the survival of the fittest individuals. The fittest strings are chosen to reproduce, crossover, and occasionally mutate, eventually evolve into a population of solutions that are highly adapted to the desired environment. By imitation of the natural process, GA is capable of developing solutions to the real problems. GA is being applied to a variety of problems and becoming an important tool in machine learning and function optimization. A simple GA that has given good results in a variety of engineering problems is composed of three operators, i.e., selection, crossover, mutation and fitness function. A brief description of the three operators is detailed as follows.

**Selection** - The selection process copies individual strings (called parent chromosomes) into a new tentative population (known as mating pool) for genetic operations. According to Darwin's evolution theory the best ones should survive and create new offspring. There are many methods how to select the best chromosomes, for example roulette wheel selection, Boltzman selection, tournament selection, rank selection, steady state selection and some others.

**Crossover** - The crossover operator oversees the mating process of two chromosomes. Two parent chromosomes are selected from the mating pool randomly and the crossover rate, which is a real number between zero and one, determines the probability of producing a pair new chromosome from the parents, which are called the offspring. Offspring inherits complementing genetic material from its parents. Some of the commonly used techniques are one-point crossover, two-point crossover and multipoint crossover.

**Mutation** - The main aim of mutation is to introduce genetic diversity into the population. Sometimes, it helps to regain the information lost in earlier generations. Like natural genetic systems, mutation in GA is made occasionally. A random position of a random string is selected and replaced by another character from the alphabet. In the case of binary representation it negates the bit value and is known as bit mutation. Normally, mutation rate is kept fixed in order to sustain diversity, (which may be lost due to crossover and very low mutation rate) into the population. Mutation is not always worth performing. High mutation rate can lead to genetic search to a random one. It may change the value of an important bit and thereby affect the fast convergence to a good solution. Moreover, it may slow down the process of convergence at the final stage of GA.

**Fitness function** - In order to evaluate how good the different individuals in the population are, a fitness function needs to be defined. In achieving the objectives to minimize the error of the predicted output with the targeted output, and to encourage the inclusion of Don’t Care (DC) features that minimizing the number of input feature, GA fitness function employed is defined as \( \text{fitness} = pf_1 + qf_2 \), where \( f_1 \) is the mean square error of prediction, \( f_2 \) is the percentage of features that has been assigned as DC (i.e., compression rate of input feature), \( p \) and \( q \) are user select parameters.
3. Experiments and Results
The NOx dataset consists of 3405 data samples. Each of the data consists of 155 input features that take from the parameter of the plant. Meanwhile, there is one single target output which represents the concentration of the NOx. The dataset are divided into three subsets; 50% for training, 25% for prediction and 25% for test. In this experiment, the GA settings are
- Population Type = 50
- Population Size = 150
- Generation Limit = 500

First, the dataset is tested using LR method, the mean square error for validation and test sets are 0.063 and 0.144 respectively. These results are used as benchmarks for comparison with results of GA-LR.

In second stages, the GA-LR is used to learn the training set with different values of user select parameter $p$ and $q$. The respective mean square errors of validation set as well as number of feature used are recorded in Table 1. The smallest error is 0.028 (when $p = 1.00$ and $q = 0.00$) with 80 features. The features are not much reduced compared to the acceptable error, 0.058 (when $p = 0.80$ and $q = 0.15$) with only 11 features. After the parameters are determined ($p = 0.80$ and $q = 0.15$), GA-LR algorithm predicts that the error of the testing data is 0.133.

In third stage, further experiments have been conducted the test the sensitivity of the performance of GA-LR with different setting (numbers of population and generation) of GA. Table 2 indicates that GA-LR is insensitive to the change of GA’s parameters setting. The difference of the features selected and the errors are rather small.

**Table 1.** Mean square errors and number of features selected (in parenthesis) of validation set based on $p$ and $q$

| $p$ | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|-----|-----|-----|-----|-----|-----|-----|
| 0.00 | 0.045 (18) | 0.053 (14) | 0.051 (14) | 0.041 (20) | 0.038 (23) | **0.028 (80)** |
| 0.05 | 0.053 (18) | 0.057 (15) | 0.052 (23) | 0.049 (16) | 0.049 (16) | 0.034 (89) |
| 0.10 | 0.060 (17) | 0.054 (17) | 0.055 (21) | 0.052 (17) | 0.048 (18) | 0.032 (83) |
| 0.15 | 0.059 (16) | 0.059 (15) | 0.058 (14) | **0.058 (11)** | 0.050 (25) | 0.043 (84) |
| 0.20 | 0.060 (12) | 0.058 (14) | 0.060 (12) | 0.059 (17) | 0.048 (27) | 0.053 (88) |
| 0.25 | 0.058 (13) | 0.055 (14) | 0.055 (15) | 0.057 (19) | 0.050 (22) | 0.037 (83) |
| 0.30 | 0.059 (16) | 0.058 (17) | 0.055 (18) | 0.055 (17) | 0.050 (27) | 0.042 (90) |

**Table 2.** Results of GA-LR for different GA’s parameters

| Generation | Population | Number of Selected Features | Mean Square Error |
|------------|------------|------------------------------|-------------------|
|            |            |                              | Train | Testing |
| 150        | 50         | 11                           | 0.058 | 0.113   |
|            | 100        | 8                            | 0.060 | 0.111   |
|            | 150        | 7                            | 0.062 | 0.135   |
| 300        | 50         | 8                            | 0.060 | 0.141   |
|            | 150        | 8                            | 0.061 | 0.132   |
|            | 250        | 9                            | 0.056 | 0.108   |
| 600        | 250        | 7                            | 0.062 | 0.123   |
|            | 450        | 7                            | 0.062 | 0.123   |
|            | 600        | 7                            | 0.062 | 0.123   |
4. Summary
In this research, a new algorithm (i.e., GA-LR) has been proposed which perform prediction to the NOx emission in power generation plant. The GA-LR is conceived to be a better NOx emission predictor with the least features selection while maintaining the lowest possible error. The proposed GA-LR allows reliable prediction of NOx emission which can help the monitoring of the gas pollution. Early action can be taken into measure to reduce the gas pollution from the power generation plant. Further works can be focused on other type of gases and improving the error by different methods.

References

[1] Azid I. A., Ripin Z. M., Aris M. S., Ahmad A. L., Seetharamu K. N. and Yusoff R. M. 2000 Predicting combined-cycle natural gas power plant emission by using artificial neural networks in Proc. Of IEEE TENCON (Kuala Lumpur, Malaysia, Sep. 2000) 512-517

[2] Tronci S., Baratti R. and Servida A. 2002 Monitoring pollutant emissions in a 4.8MW power plant through neural networks Neurocomputing (Mar. 2002) 43 3-15

[3] Zheng L., Zhou H., Wang C. and Cen K. 2008 Combining Support Vector Regression and Ant Colony Optimization to Reduce NOx Emissions in Coal-Fired Utility Boilers ACS Publications: Energy Fuels 22 1034-1040

[4] Yang J. and Honavar V. 1998 Feature Subset Selection Using A Genetic Algorithm IEEE Intelligent Systems and Their Application 13 44-49

[5] Jain A. and Zonker D. 1997 Feature Selection: Evaluation, Application, and Small Sample Performance IEEE trans. on Pattern Analysis and Machine Intelligence 19 153-158

[6] Neter J., Kurtnet M., Nachtsheim C. J. and Wasserman W. Applied Linear Regression Model (McGraw Hill/Irwin Series Operations and Decision Sciences)

[7] Afifi A. A. and Elashoff R. M. 2012 Missing Observation in Multivariate Statistic II, Point Estimation in Simple Linear Regression Journal of American Statistical Association 62 10-29

[8] Fang Y., Hans-George M. and Jane-Ling W. 2005 Functional Linear Regression Analysis for Longitudal Data Project Euclids: Annals of Statistics (vol 33) 2873-2903

[9] Huang G. B., Wang D. H., and Lan Y. 2011 Extreme Learning Machines: A Survey, International Journal of Machine Leaining and Cybernetics 2 107-122

[10] Wei H., Tong Z., and Jian S. 2004 Use of a genetic algorithm to optimize multistage erbium-doped fiber-amplifier systems with complex structures US National Library of Medicine National Institutes of Health 12 531–544