A COMPREHENSIVE STUDY ON DEVELOPING NEURAL NETWORK MODELS FOR PREDICTING THE COAGULANT DOSAGE AND TREATED WATER QUALITIES FOR A WATER TREATMENT PLANT

by

CHAMANTHI DENISHA JAYAWEERA

Thesis submitted in fulfillment of

the requirements for the degree of

Master of Science

July 2019

ACKNOWLEDGEMENT

I wish to express my profound gratitude towards my supervisor, Associate Professor Dr. Norashid Aziz, for guiding me through the research with patience and kindness. I am also thankful for the assistance provided by the Sabah Water Supply Department and LDWS for providing data of the Segama Water Treatment Plant in Lahad Datu, Sabah. I would like to acknowledge Mr. Asabri Robenson for his effort in acquiring the data used for this research from the Sabah Water Supply Department and LDWS.

I am grateful for the support and financial assistance provided by my parents, B.M.P.S. Kumarihamy and H.C. Jayaweera, to complete my Master’s degree. Finally, I thank the School of Chemical Engineering of Universiti Sains Malaysia, for providing me with the opportunity to carry out the research.

Chamanthi Denisha Jayaweera
15th April 2019

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial neural network
BMU	Best matching unit
ELM	Extreme learning machine
ELM-RBF	Extreme learning machine radial basis function neural network
ELM-SLFN	Extreme leaning machine single layer feed forward neural network
GA	Genetic algorithm
GRNN	Generalized regression neural network
GRR	Gaussian reference rule
MLP	Multilayer perceptron
MSE	Mean squared error
OLS	Orthogonal least squares method
PCA	Principal Component Analysis
PCC	Pearson correlation coefficient
RW/TW	Raw water/ Treated water
SOM	Self-organizing map
TDS	Total dissolved solids
TOC	Total organic carbon
UV	Ultra violet

SATU KAJIAN KOMPREHENSIF TERHADAP PEMBANGUNAN MODEL RANGKAIAN NEURAL UNTUK MERAMAL DOS BAHAN PENGGUMPAL DAN KUALITI AIR TERAWAT DALAM LOJI PERAWATAN AIR

ABSTRAK

Penentuan dos bahan penggumpal optimum untuk rawatan air secara tradisinya dijalankan menggunakan ujian balang yang merupakan prosedur memakan masa dan tidak berupaya untuk bertindak balas terhadap perubahan mendadak dalam kualiti air. Oleh itu, teknik pemodelan didorong data seperti rangkaian neural digunakan untuk membangunkan model ramalan untuk proses penggumpalan. Dalam kerja ini, tiga rangkaian rangkaian neural yang berbeza iaitu rangkaian neural regresi umum (GRNN), rangkaian neural lapisan tunggal suap depan dengan mesin pembelajaran melampau (ELM-SLFN) dan rangkaian neural asas jejarian dengan mesin pembelajaran melampau (ELM-RBF) telah dibangunkan untuk meramalkan dos bahan penggumpal dan prestasi mereka dibandingkan dengan rangkaian neural perseptron berbilang lapis (MLP) yang biasa digunakan. Ia menunjukkan bahawa model ELM dan GRNN menggunakan usaha dan masa yang lebih rendah untuk latihan berbanding dengan MLP. ELM-RBF menunjukkan keseimbangan terbaik antara ketepatan ramalan dan keperluan pengiraan. Oleh itu, ELM-RBF telah digunakan untuk membangunkan model untuk meramalkan dos bahan penggumpal, kekeruhan air terawat (TW) dan sisa aluminium dengan nilai R masing-masing 0.9752, 0.8239 dan 0.9019. Parameter input yang diperlukan untuk membangunkan setiap model ditentukan dengan menggunakan algoritma carian menyeluruh global kerana ia telah ditunjukkan bahawa pekali korelasi Pearson dan analisis komponen utama merupakan teknik yang tidak sesuai untuk memilih parameter input untuk bagi kajian ini. Oleh itu, input yang digunakan untuk meramalkan dos bahan penggumpal ialah kekeruhan air mentah (RW), warna RW dan alum (t-1). Keberkesanan dos bahan penggumpal dan model kualiti TW telah dipertingkatkan dengan menggunakan model imputasi dan algoritma genetik. Model imputasi telah dibangunkan menggunakan kaedah K-kelompok dengan ketepatan imputasi yang serupa dengan peta swaorganisasi, untuk menangani kegagalan perkakasan sensor yang menyebabkan masa henti dalam loji rawatan air automatik dan untuk memastikan penggunaan model dos bahan penggumpal yang berterusan. Nilai yang hilang dari kekeruhan RW dan warna RW dibina semula menggunakan model imputasi dengan nilai R masing-masing 0.9075 dan 0.8250. Selepas itu, kekeruhan RW dan warna RW yang dibina semula digunakan untuk meramalkan dos bahan penggumpal dengan nilai R 0.9742 dan 0.9809 adalah sangat memuaskan. Manakala GA menaikkan nilai R daripada model kekeruhan TW kepada 0.8294. GA meningkatkan keupayaan ELM-RBF untuk mengenal pasti tindak balas yang diperlukan dari kekeruhan TW terhadap dos alum.

A COMPREHENSIVE STUDY ON DEVELOPING NEURAL NETWORK MODELS FOR PREDICTING THE COAGULANT DOSAGE AND TREATED WATER QULAITIES FOR A WATER TREATMENT PLANT

ABSTRACT

Determination of the optimum coagulant dosage for water treatment is traditionally carried out using the jar test, which is a time consuming procedure incapable of responding to sudden changes in water qualities. Therefore, data driven modeling techniques such as neural networks are used for developing predictive models for the coagulation process. In this work, three different neural network models, namely, the general regression neural network (GRNN), extreme learning machine single layer feed forward neural network (ELM-SLFN) and the extreme learning machine radial basis function neural network (ELM-RBF) were developed to predict the coagulant dosage, and their performances were compared with the commonly used multilayer perceptron neural network (MLP). It was shown that the ELM and the GRNN models consumed significantly lesser effort and time for training compared to the MLP. The ELM-RBF demonstrated the best tradeoff between prediction accuracy and computational requirement. Therefore, the ELM-RBF was used to develop models for predicting the coagulant dosage, treated water (TW) turbidity and residual aluminum with R values of 0.9752, 0.8239 and 0.9019 respectively. The input parameters required to develop each model was determined using a global exhaustive search algorithm as it was shown that the Pearson correlation coefficient and the principal component analysis were not suitable techniques for selecting input parameters for this study. Thus, inputs used for predicting the coagulant dosage were raw water (RW) turbidity, RW color and alum (t-1). The effectiveness of the coagulant dosage and the TW quality models were improved using an imputation model and a genetic algorithm. The imputation model was developed using K-means clustering with an imputation accuracy similar to a self-organizing map, to cope with failures in hardware sensors causing downtime in fully automated water treatment plants and ensure the continual use of the coagulant dosage model. The imputation model reconstructed missing values of RW turbidity and RW color with R values of 0.9075 and 0.8250 respectively. Subsequently, the reconstructed RW turbidity and RW color were used to predict the coagulant dosage with R values of 0.9742 and 0.9809 respectively, which are highly satisfactory. Meanwhile, the GA improved the R value of the TW turbidity model to 0.8294. The GA significantly improved the ability of the ELM-RBF to identify the required response of TW turbidity to the alum dosage.

CHAPTER ONE

INTRODUCTION

1.1 Project background

Water treatment is carried out by physical separation of solid pollutants and addition of measured dosages of chemicals. Maintenance of a water treatment plant requires regular determination of operational variables such as flow rates, temperatures, aeration rate and required concentration of chemicals. Optimal values for such operational variables, depending upon the raw water parameters, are determined via routine experiments, which cost chemicals, time and capital.

The coagulation process is a vital stage of water treatment, where minute particulate substances, which are responsible for the turbidity and color of raw water, are accumulated to form larger removable agglomerates. This is accomplished by the addition of coagulant, the dosage of which should be precisely measured in a regular basis to ensure the efficient removal of the aforementioned substances. The optimal amount of coagulant is determined by the jar test (Joo *et al.*, 2000; Yu *et al.*, 2000; Maier *et al.*, 2004), which is a reactive response to changes in treated water qualities. It is carried out by observing critical treated water parameters of a series of jars containing equal volumes of raw/waste water, which are treated with different concentrations of coagulant. Thus, jar tests regularly consume additional costly chemicals. Parameters such as coagulant types, mixing rate and aeration level/time can be optimized using the jar tests. Due to time consumed by the experiments, jar tests are incapable of responding to sudden changes of water qualities (Valentin and Denoeux, 1999).

Similarly, the Segama water treatment plant treating water from the Segama River flowing through the Lahad Datu district, Sabah, in a 24 hour basis, uses jar tests to measure the optimal coagulant dosage, and records raw water, process and treated water parameters. The primary intention of this study is to develop a model/framework that facilitates predicting optimal coagulant dosage for the Segama water treatment plant.

Linear regression, multi parameter regression, non-parametric regression and non-linear data driven techniques have been utilized to develop predictive models. As coagulation is a complex phenomenon that involves several coagulation mechanisms and factors influencing the chemistry of the process, it has been pointed out in literature that nonlinear data driven models perform better than regression models in predicting the coagulant dosage (Yu *et al*., 2000; Towler *et al*., 2009; Dharman *et al*., 2012).

Artificial neural networks (ANNs) are one of nonlinear data driven techniques that has been used to develop predictive models for the coagulation process in multiple instances (Maier *et al*., 2004; Wu and Lo, 2008 and 2010; Griffith *et al*., 2011; Dharman *et al*., 2012; Zangooei *et al*., 2016, Santos *et al*., 2017). Development of an ANN involves a series of stages that affects the performance and effectiveness of the ultimate model such as data preprocessing, input parameter selection and model development. It has been shown by Joo *et al*. (2000) that data preprocessing improves model performance by improving the learning rate and the terminal error in the procedure of neural network training. It had also improved the prediction accuracy of the test data set. Data preprocessing is carried out by outlier removal and data transformation which is commonly done through data normalization and standardization (Maier *et al*., 2004; Robenson *et al*., 2009; Dharman *et al*., 2012; Deng and Lin, 2016). The subsequent stage of model development is input parameter selection which could be carried out using techniques such as Pearson correlation coefficient, principal component analysis, cross correlation and forward regression.

ANNs can be categorized based on the model architecture such as radial basis function neural networks (RBF), general regression neural networks (GRNN) and multilayer perceptron neural networks (MLP). MLPs are the more commonly used ANN for developing predictive models for the coagulation process. However, the best type of neural network for this study is yet to be determined based on the tradeoff between prediction accuracy and computational requirement.

1.2 Problem statement

An artificial neural network (ANN) model predicting the coagulant dosage is developed as an effective replacement to the jar test, which is a time consuming procedure incapable of responding to sudden changes in water qualities (Maier et al., 2004; Robenson et al., 2009; Santos et al., 2017). The most commonly used type of neural network is the multilayer perceptron neural network (MLP). Development of an MLP requires optimizing several model parameters such as the number hidden layers, number of hidden neurons in each layer, activation functions of each layer, the learning rate and momentum term. Due to the iterative learning algorithms used in the MLP, the training process alone consumes a long time. As optimizing each of the above mentioned parameters requires training the model with every modification on each model parameter, the long training time is multiplied by a large factor. Therefore, a large effort has to be incurred in developing an MLP. Researchers have implemented genetic algorithms on the MLP to train the model such that it produces solutions at the global optimum (Bagheri *et al*., 2015; Wibowo *et al*., 2018; Abdollahi *et al*., 2019), where the MLP still has to be iteratively trained several times to reach the minimum error solution. Therefore, implementing additional model improvement measures for the MLP requires a significant additional effort due to the time consumed by the iterative training procedure of MLPs.

The general regression neural network (GRNN) is an alternative ANN that can be used for addressing the shortcomings of the MLP (Kennedy *et al*., 2015; Kim and Parnichkun, 2016). However, there have been instances where the prediction accuracy of GRNN was less than the MLP (Kim and Parnichkun, 2016). Studies carried out on developing predictive models for the coagulation process have simply used the training data set as radial basis centers and determined the smoothing parameter in a subjective manner (Heddam et al., 2011; Kennedy et al., 2015; Kim and Parnichkun, 2016), which may have affected the performance of the GRNN. Although it has been generally stated that using K-means clustering improves the GRNN, no study has demonstrated how effectively clustering techniques could improve the GRNN in coagulation applications. Additionally, center selection techniques such as the orthogonal least squares algorithm have not been implemented on the GRNN. Techniques for determining the smoothing parameter such as the Gaussian reference rule are yet to be implemented on GRNNs developed for aiding the coagulation process. Therefore, it is difficult to select the most suitable center selection technique and smoothing parameter determination technique for coagulation modeling from literature.

The extreme learning machine (ELM) neural networks were introduced recently by Huang et al. (2004). ELM single layer feed forward neural network (SLFN) was used for predicting the coagulant dosage by Deng and Lin (2017), where they demonstrated that the ELM-SLFN is a better alternative to the MLP, due to the less computational effort required. The only model parameter that requires adjusting is the number of hidden neurons. However, the extension of the ELM for the radial basis function (RBF) case has not been implemented for predicting the coagulant dosage. It was shown by Huang and Siew (2004) that the ELM-RBF has high generalization ability and it only needs adjusting the number of radial basis centers. Liu and Wan (2015) also proved that the ELM-RBF had a great potential of producing universally consistent solutions. Therefore, the ELM-RBF could be offering the best tradeoff between prediction accuracy and computational requirement for predicting the coagulant dosage. It is also worthy to note that a comparison between GRNN and ELM models have not been carried out for predicting the coagulant dosage. Thus, the most suitable model for this study is yet to be determined.

In addition to developing models for predicting the coagulant dosage, ANNs could also be developed for predicting treated water (TW) qualities to replace expensive hardware sensors, avoid lengthy procedures for measuring certain water qualities and for implementing an early warning system for critical water qualities that require strict control (Leardi, 2003; Maier et al., 2004). However, TW quality models should only be developed for selected parameters, in order to avoid incurring additional effort in training unnecessary models.

The selection of an appropriate technique for determining input parameters is an important aspect of model development. Most studies on coagulation modeling have employed linear techniques such as the Pearson correlation coefficient and the principal component analysis for input parameter selection (Lamrini *et al*., 2005; Wu and Lo, 2008 and 2010; Zangooie *et al,* 2016) which are incapable of capturing nonlinear forms of relationships occurring in the coagulation process between input variables and output variables (Bowden *et al*., 2005; May *et al*., 2011). However, nonlinear input parameter selection techniques are computationally intensive and are not as simple as linear techniques to be executed. Thus, the most effective input parameter selection technique needs to be determined.

When neural network models are developed for aiding the coagulation process, additional measures could be orchestrated to cope with foreseeable issues during the implementation of the model and to further improve performance if needed. One of the foreseeable issues for a fully automated coagulation process is the failure of a hardware sensor, which may cause downtime (Newhart *et al*., 2019) and reduce the effectiveness of the coagulant dosage model. Studies have been carried out to develop 'imputation models' to approximate the missing input from the failed sensor using self-organizing maps (Valentin and Denoeux, 2002; Latif and Mercier, 2010; Lamrini *et al*., 2011; Juntunen *et al*., 2013). However, training the self-organizing map (SOM) with sufficient number of map units consumes time. Meanwhile, the genetic algorithm (GA) is a common technique used for improving models that are difficult to train (Whitley *et al*., 1990; Bowden *et al*., 2004; Bagheri *et al*., 2015). They are mostly implemented on MLPs for achieving the global optimum solution. Although GAs have been used to successfully improve the performance of models, GAs are considered to be time consuming due to its slow rate of convergence and the generally long training time required by MLPs. Implementing GAs on other types of ANNs such as the GRNN and ELM-ANNs (Qinfeng, 2016) are bound to consume less time as their training processes are not iterative. However, the use of genetic operators that unnecessarily cause the algorithm to converge slowly may reduce the advantage of using GRNNs/ELM ANNs. Thus, the aforementioned additional measures which are developed for increasing the effectiveness of the models may end up reducing the effectiveness of the models due to their time consuming nature.

1.3 Research objectives

The objectives of this work are as follows.

1. To test the reliability of commonly used linear input parameter selection techniques in developing predictive models for the Segama water treatment plant.
2. To determine the best neural network model among the MLP, GRNN, ELM-SLFN and ELM-RBF in predicting the coagulant dosage, TW turbidity and residual aluminum.
3. To improve the effectiveness of the neural network models developed using an imputation model and a genetic algorithm with minimal effort.