APTIKOM Journal on Computer Science and Information Technologies
Vol. 3, No. 3, 2018, pp. 84~91
ISSN: 2528-2417
DOI: 10.11591/APTIKOM.J.CSIT.125

Genetic algorithm artificial neural network in near infrared spectroscopic quantification

Hasan Ali Gamal Al-Kaf, Kim Seng Chia*, Nayef Abdulwahab Alduais, Musaed Al-Subari
Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia,
Parit Raja, Johor, Malaysia.
*Corresponding author, kschia@uthm.edu.my

Abstract

The implantation of a genetic algorithm (GA) in quantitating components of interest in near infrared spectroscopic analysis could improve the predictive ability of a regression model. Thus, this study investigates the feasibility of a single layer Artificial Neuron Network (ANN) that trained with Levenberg-Marquardt (SLM) coupled with GA in predicting the boiling point of diesel fuel and the blood hemoglobin using near infrared spectral data. The proposed model was compared with a well-known model of Partial Least Squares (PLS) with and without Genetic Algorithm. Results show that the proposed model achieved the best results with root mean square error of prediction (RMSEP) of 3.6734 and correlation coefficient of 0.9903 for the boiling point, and RMSEP of 0.2349 and correlation coefficient of 0.9874 among PLS with and without GA, and SLM without GA. Findings suggest that the proposed SLM-GA is insusceptible to the number of iterations when the SLM was trained with excessive iteration after the optimal iteration number. This indicates that the proposed model is capable of avoiding overfitting issue that due to excessive training iteration.

Keywords*: genetic algorithm, near infrared spectroscopy, neural network, partial least square, single layer*

1. Introduction

Near infrared (NIR) spectroscopy has gained its popularity as a superior analytical tool due to its potential in various fields' analytical applications e.g. food, agriculture, medicine, petroleum, petrochemical environmental, clinical, biological, and biomedical sectors during the past few years. This revolutionary technique has become widely accepted because of its benefits in acquiring spectra for solid and liquid samples rapidly and non-invasively, without a conventional prior sampling process that is time consuming and expensive. These features make NIR attractive for speedy and straightforward natural and synthetic products characterization. Due to the complexity of the NIR data, data dimensional reduction techniques help in providing model accuracy by eliminating redundant and over-lapping information. In conjunction with variable selection or latent variable extractions, the complexity of the NIR data would be reduced while the robustness of the model against disturbance would be increased which in return make model interpretation much easier [1]. Several data dimensional reduction methods have been proposed to evaluate the effects of wavelength selections. Among various approaches, Genetic Algorithm (GA) coupled with Partial Least Squares (PLS) appears to be one of popular directions. This could be due to the capability of GA that is robust with the goal of search that mimics the principles of natural selection to solve large optimization problems [2-4], while PLS is a popular predictive model in NIR spectroscopic analysis. For instance, previous works report that Genetic Algorithm-based Wavelength Selection (GAWS) achieved better than PLS in two real datasets [5], GA that coupled with PLS (GA-PLS) improves the performance of soil type discrimination [6] and the visualization of Pseudomonas loads in chicken fillets [7]. Additionally, there are various approaches to adapt GA. For example, GA has been proposed to use only a mutation operator that achieved better performance compared with successive projection algorithms, PLS and classical formulation of GA [8]. A combination of adaptive boosting algorithm and GA for PLS in selecting the most informative wavelengths in NIR spectroscopy has been reported to achieve better prediction accuracy compared to that without the adaptive boosting algorithm [9].

Besides GA-PLS, several studies have been carried out to investigate the GA that coupled with artificial neural network (GA-ANN). This is because the GA has a potential to solve the problems of back propagation algorithm with the steepest descent algorithm which has a major limitation in the solutions getting trapped at local minima. For instance, evolutionary GA could substitute the backpropagation

Received July 10, 2018; Revised October 3, 2018; Accepted October 15, 2018

algorithm to look for global minima at various discrete locations in a huge space simultaneously. Additionally, GA could be used to optimize the ANN network architecture and reduce the learning time through avoiding conventional iterative method and minimizing error systematically [10-11].

Recently, single layer ANN that trained with Levenberg-Marquardt (LM) model has been found outperformed to multilayer layer ANN that trained with LM, PLS, and extreme learning machine in predicting the diesel fuel properties [12]. The performance of this model that has a simple structure and one tunable parameter may be further improved using GA. This is because a further reduction of the data complexity by means of GA might ultimately improve the robustness and accuracy of this model. Thus, this study aims to investigate the feasibility of a single layer ANN that trained with LM coupled with GA in predicting the petroleum oil boiling point and blood hemoglobin in NIR spectroscopic analysis, and to compare the proposed model with PLS-GA [13].

2. Material and Method

2.1. Dataset

Based on the measured results of NIR spectra of boiling point at Southwest Research Institute from a project sponsored by the US army which composes the diesel fuel dataset. Eigenvector research homepage at http://www.eigenvector.com provides the dataset. The samples were divided into three sets with names and extension of letter b, a, and h in according to datasets grouping in the website, respectively, for calibration, validation, and prediction sets. The second dataset was provided by Karl Norris from IDRC shootout 2010 [14]. Blood samples were analyzed with a NIRSystems 6500 spectrometer. All spectra have 700 variables, from 1100 to 2498 nm, with a 2 nm interval the blood hemoglobin reference was measured by a high-volume hematology analyzer, Coulter STKS monitor made by the Coulter Corporation of Hialeah, FL. The dataset contains 231 sets for calibration, 194 for validation and 58 unseen data sets for the blind test to measure predictive accuracy of the modelling networks.

2.2. Predictive Model

A single layer ANN trained with LM and PLS without variable selection were used as benchmark in predicting the petroleum oil boiling point and blood Hemoglobin using NIR data. The model analysis contains two phases namely; training phase and prediction phase. First, the dataset of boiling point and hemoglobin were randomly separated into training and testing datasets. Both dataset were analyzed in the previous studies [12,15]. The same training dataset were used to train predictive models while the predictive accuracy of the optimal model would be evaluated using the testing data set in the second phase. The root means square error and the regression coefficient of the model are used to evaluate the model.

2.3. Genetic Algorithm

Over the past decades, GA has gained its popularity as an optimization technique which employs probabilistic, non-local search process. Selection of variables for multivariate calibration is considered as optimizing the process. Besides that, a genetic algorithm is used to implement an automated wavelength selection procedure for use in construction multivariate calibration and selecting effective wavelengths. Furthermore, GA is capable of calibrating mixtures with almost identical spectra without loss of prediction capacity. Figure 1 shows the flowchart of the GA implementation of the single layer ANN trained with LM and PLS. This process starts with the preparation of training, validation, and testing dataset. Then, the wavelength selection is implemented in two ways briefly; the wavelength automated-optimal selection and the predetermined wavelength selection such as 15, 50 and 200. After that, a set of real and binary values are selected and implemented to produce a chromosomic solution. Subsequently, the random initial population of either 100 or 400 is selected. Then, the response associated with the corresponding experimental conditions is evaluated for each chromosome. This evaluation is achieved through evaluating the root mean square error (RMSE) for each chromosome.

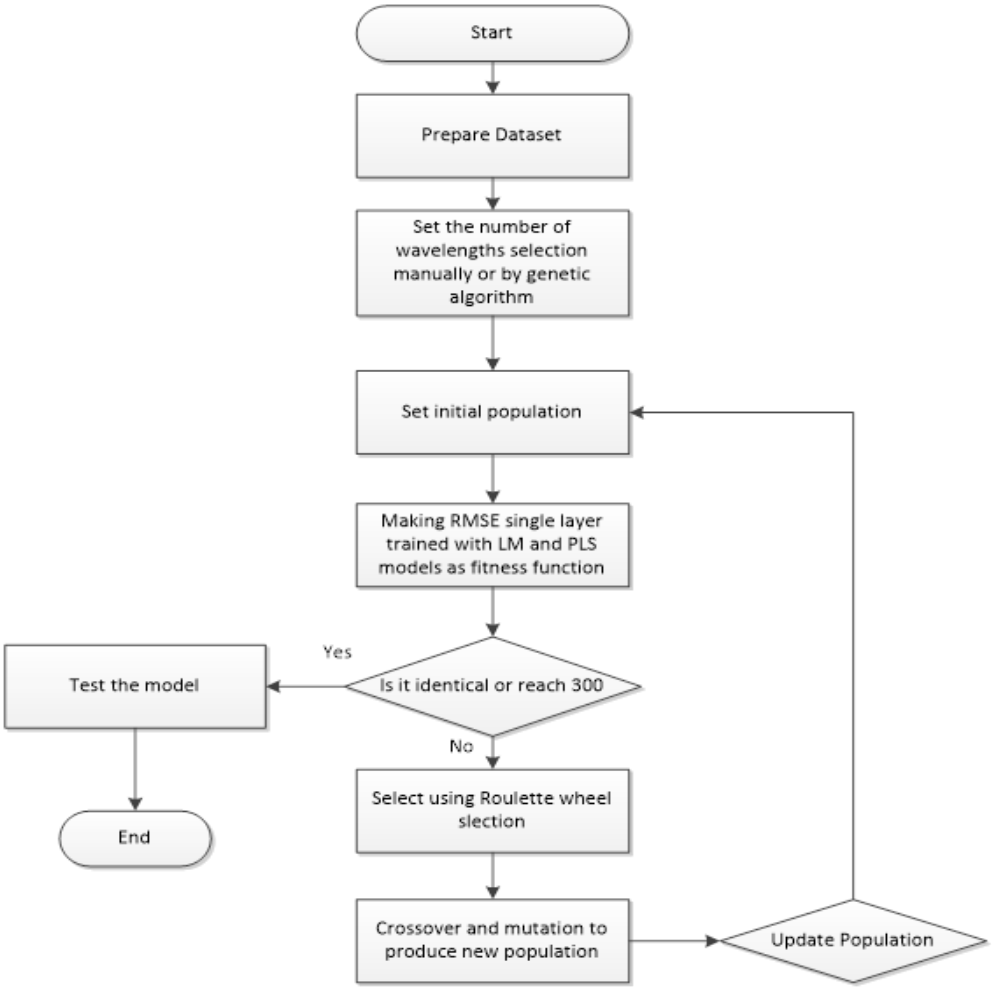

Figure 1. The proposed Genetic Algorithm for ANN and PLS

In accordance with that, Roulette selection arranges the chromosomes' and its rank in descending to fitness value accordingly. Then, the new population that can be considered as the next generation created by the reproduction step which is made up by the uniform crossover of the original chromosomes. The chromosomes with a high fitness value have a higher probability to reproduce than others chromosome with the target to improve the overall fitness of the population. However, there are some problems that might occur, which can be overcome by mutations. The essential problem to be solved occurs if a variable is not selected from any of the original chromosomes, as it would never be selected in the coming generations if mutations were not present. A mutation is simply an inversion of a gene in a chromosome with a mutation rate in between 0.001 and 0.01. After the termination condition is fulfilled which when data is identical or reach 300 iterations. Then best wavelength is interring into predictive models to calculate RMSEP and correlation coefficient for prediction in a first way, and in a second way after the generation is stopped, then it validated using RMSEV and tested using RMSEP. All the result are obtained for only one run. The fitness function is the root mean square error (RMSE). To construct fitness functions; two methods are adopted. For the first method, the fitness function is trained using RMSEC once the iterative GA process is finished which latterly is used to validates using RMSEV and tested using RMSEP. While the second method, the fitness function is validated using RMSEV after the iterative process of the selected wavelength is conducted. Then the performance of the model is tested using RMSEP.

2.4. Parameter Setting for Single Layer ANN That Trained with GA

The general default training parameters for single layer ANN that trained with LM was same as Matlab default parameter setting. The two public available source codes [16-17] were adopted and successfully modified to implement GA in this study. The first source code was used when the automated

wavelength was selected by GA. On the other hand, the second source code was used when the predetermined wavelength was chosen intuitively. Settings of the GA are dividing into two parts. Firstly, the parameters setting for a predetermined selected wavelength method are as follows, the initial population was 100, the population type was real, the selection method was Rouleteweel, Crossover method was uniform crossover, and the number of generation was 300. This process was repeated using the initial population of 400 to evaluate the effect of the initial population. The parameters setting for automated wavelengths selection are as follow, the initial population was 100, population type was binary, the selection method was Rouleteweel, crossover method was arithmetic, and the number of generation was 300. To evaluate the predetermined wavelength using the proposed model SLM-GA for predicting the boiling point and hemoglobin, RMSEV was used as the fitness function. The predetermined wavelengths selection were 15, 50 and 200 for both datasets. The initial population were 100. The generation was set to 300 or the result was taken when the result was identical. The process was repeated using the initial population of 400.

3. Results and Discussion

3.1. Parameters Optimization for PLS and ANN

Figure 2 illustrates the performance of PLS in terms of RMSEP and RMSEC for both petroleum oil boiling point and blood hemoglobin datasets. The determination of the optimal factor number is crucial, in which after an optimal number of PLS's factors, the value of root mean square error goes high that indicated an overfitted function. The optimal number of PLS factors for boiling point and hemoglobin datasets were 15 and 19, respectively. Figure 3 shows the performance of the single layer ANN that trained with LM in terms of RMSEP and RMSEC for both petroleum oil boiling point and blood hemoglobin datasets. The optimal iterations numbers are 11 for boiling point and 9 for hemoglobin. After a certain number of iterations, the values of RMSEP and RMSEC shows dramatic changes which indicting an overfitting of boiling point and hemoglobin datasets at 7 and 6 respectively.

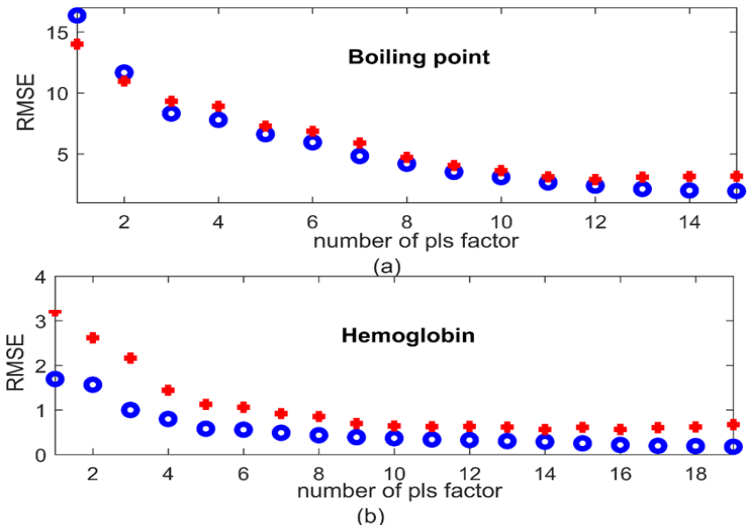

Figure 2. Optimal factor for PLS for boiling point and hemoglobin where + for RMSEP and $_{O}$ for RMSEC: (a) boiling point (b) hemoglobin dataset

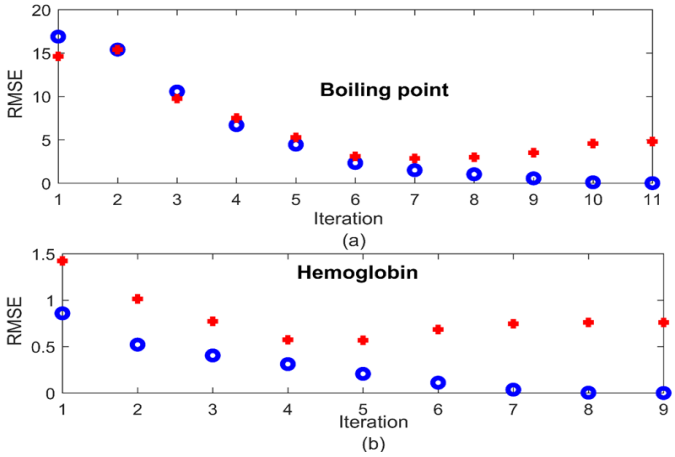

Figure 3. Optimal iterations for single layer for boiling point and hemoglobin where + for RMSEP and $_{O}$ for RMSEC: (a) boiling point, (b) hemoglobin dataset

3.2. Comparisons between single layer ANN and PLS coupled with GA

In order to compare the performance between single layer ANN trained LM with genetic algorithm (SLM-GA) and PLS with GA (PLS-GA), the simulation set is, the population size sets to 100 whereas two different fitness functions were used as explained in methodology. The optimal results were achieved when the number of iterations was identical or reaches 300. Table 1 shows the simulation results of the proposed SLM-GA algorithm compared to PLS-GA algorithm for predicting the boiling point of diesel fuel and hemoglobin. For boiling point dataset, using RMSEC as the fitness function. It shows clearly that single layer trained with LM coupled with GA outperforms PLS with a GA which has lower RMSEP and higher correlation coefficient. Nevertheless, the model that used full spectrum performs better than SLM-GA, this is because using RMSEC as fitness function lead to overfitting while PLS-GA outperforms better than that used full spectrum. However, when RMSEV is used as the fitness function. SLM-GA perform better result than PLS-GA and full spectrum with RMSEP of 3.6, correlation coefficient of 0.9903, and wavelengths of 157 were chosen compared than 400 wavelengths for the full spectrum. For hemoglobin, SLM-GA outperforms PLS-GA and full data set. The RMSEC and RMSEP for SLM-GA were 0.2242 and 0.2349, compared than PLS-GA of 0.2537 and 0.2438, respectively. While for the correlation coefficient of calibration and prediction for SLM-GA were 0.9890 and 0.9874, compared than PLS-GA of 0.9853 and 0.9868, respectively. The SLM-GA select 259 with RMSEP of 0.2349 compare than 700 wavelengths with RMSEP of 0.2408 for that used full spectrum. PLS-GA selects 228 and has higher accuracy than that used full spectrum with RMSEP of 0.2853 and 700 wavelengths

Table 1. Comparison between single layer and partial least square coupled with genetic algorithm

Regression model	Dataset	Fitness function	Wavelength selection	Optimal iteration/ factor	RMSEC	RC	RMSEV	R_V	RMSEP	R_P
SLM-GA	Boiling point	RMSEC	284	7	1.4126	0.9965	2.8719	0.9818	4.3902	0.9856
		RMSEC	267	7	14015	0.9966	2.8689	0.9819	4.4059	0.9854
		RMSEP	157	7	2.0162	0.9931	2.0945	0.9901	3.6734	0.9903
	Hemoglobin	RMSEC	500	6	0.2044	0.9928	0.5611	0.9864	0.2242	0.9890
		RMSEP	259	6	0.2507	0.9891	0.3783	0.9915	0.2349	0.9874
PLS-GA	Boiling point	RMSEC	147	12	1.5193	0.9960	3.0972	0.9787	4.7470	0.9834
		RMSEP	169	12	2.3559	0.9904	2.1494	0.9897	4.2635	0.9860
	Hemoglobin	RMSEC	292	14	0.1697	0.9950	0.5934	0.9826	0.2537	0.9853
		RMSEP	228	14	0.2291	0.9909	0.3104	0.9940	0.2438	0.9868
PLS	Boiling point	-	Full spectrum	14	2.3878	0.9901	2.8877	0.9819	5.0975	0.9850
SLM	Boiling point	-	Full spectrum	7	1.4898	0.9962	2.8482	0.9923	4.2219	0.9881
SLM	Hemoglobin	-	Full spectrum	6	0.2063	0.9926	0.5673	0.9853	0.2408	0.9868
PLS	Hemoglobin		Full spectrum	14	0.3234	0.9818	0.6302	0.9818	0.2853	0.9819

Figure 4 show the performance of RMSEC and RMSEV of SLM-GA and PLS-GA algorithms for boiling point and hemoglobin, respectively. The result shows that lower considerable value achieved by RMSEC when RMSEC was used as the fitness function compared to that when RMSEV was used as a fitness function for both datasets. However, RMSEP suggests that this approach causes overfitting in some cases by unseen dataset. Thus, using RMSEV as the fitness function is recommended to avoid overfitting and produce a stable result.

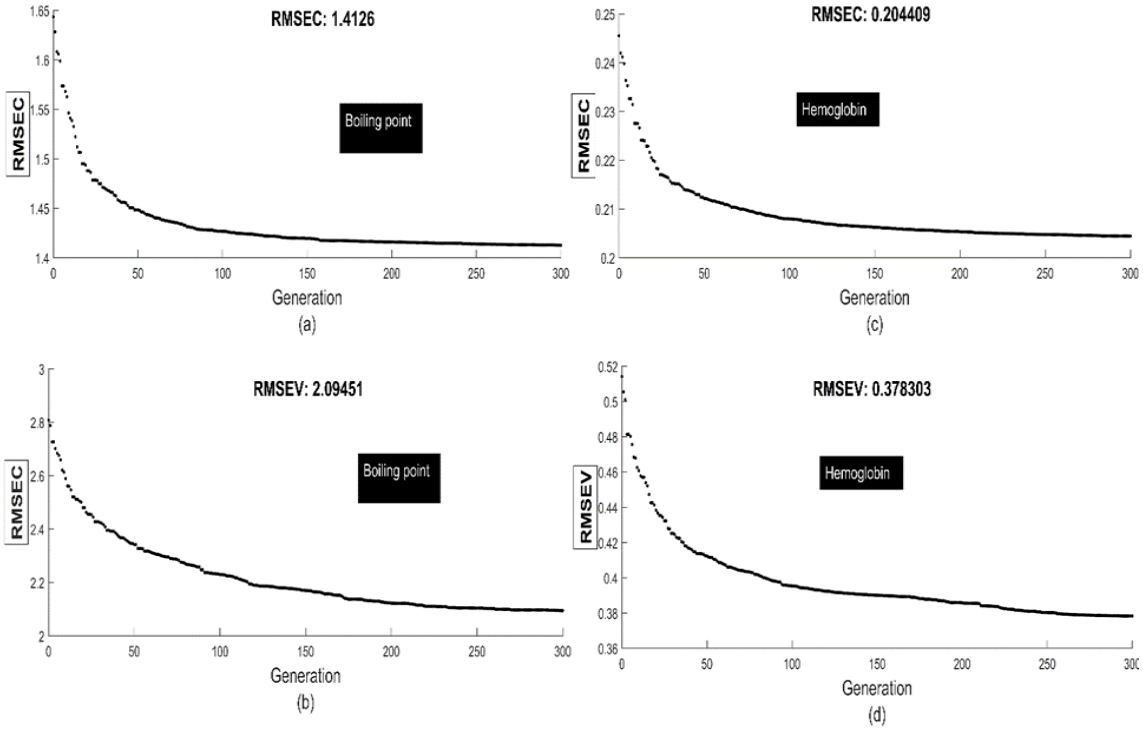

Figure 4. The performance of SLM-GA for boiling point and hemoglobin where (a) RMSEC as a fitness function for boiling point dataset (b) RMSEV as a fitness function for boiling point dataset (c) RMSEC as fitness function for hemoglobin dataset (d) RMSEV as a fitness function for hemoglobin dataset

The single layer ANN and PLS have only one parameter need to optimize i.e. optimal iteration and optimal factor, respectively. Moreover, it is recommending to use RMSEV as the fitness function to avoid overfitting. The increasing of the wavelengths number is not proportional to the improve RMSEP. Thus, choosing optimal wavelengths could help produce better accuracy. The superiority of SLM and PLS that have the ability to produce fix result that helps to identify the optimal training iteration in which the result is identical and does not produce worse result during the running of generations. The proposed model outperforms PLS-GA could be due to the following reasons. First, the LM algorithm improved the accuracy of a single layer ANN by keeping the simplicity of the model as that stated in the previous study [12]. Moreover, iteration optimization has further improved the way of the LM algorithm to perform better result with the ability to insusceptible the iteration number after its optimal iteration.

3.3. Evaluation of Predetermined Wavelength Using Genetic Algorithm

In a genetic algorithm, the population size and number of generation affect the performance of the genetic algorithm, for example, lower population size may reduce the accuracy of the model while choosing bigger population size may increase the computational time and improve the accuracy of the model. Moreover, it is important to determine the suitable number of the generation which choosing lower number of generation may not reach to the optimal result and choose a higher number of generation overfit the result. Therefore, a new parameter which is predetermined wavelength selection should be considered for

investigation in near infrared data which is not considered as a big data. In this study, we investigate the feasibility of select fix wavelength in order to improve the accuracy of the model and select a low number of wavelengths which is vital in some application to reduce the cost. Evaluating genetic algorithm of single layer trained with LM using predetermined wavelengths is shown in Table 2 which shows the boiling point and hemoglobin dataset. The selected wavelengths are 15, 50 and 200 whereas 100 and 400 are two different population sizes. RMSEV is used as a fitness function for both datasets whereas the generation is set for 300 generation and the values are taken before the over-fit occurred.

Comparing and analyzing the initial population and selected wavelength for boiling point, it is observable that, both 100 and 400 initial population don't produce better RMSEP than full data set when only 15 wavelength selection is used which showed clearly that there are many vital wavelengths are missing, as a result, reducing the accuracy of the model. For 50 wavelength selection, it produces better RMSEP than that used full spectrum when 100 initial population is used and approximate RMSEP when 400 initial population is used. Finally, for 200 wavelength selection, it outperforms that used full spectrum when 200 initial population is used but lowers RMSEP when 100 initial population is used.

For hemoglobin datasets, 200 wavelengths for both of the selected initial populations outweigh the predicted value compares to other selected wavelengths and full spectrum. Furthermore, 50 wavelengths have approximately similar results to that used full spectrum when 400 initial population is chosen. Obviously, 15 wavelengths produce higher RMSEP for both of selected population size.

Table 2. Predetermined wavelengths selected by single layer trained coupled with genetic algorithm

Dataset	Wavelength selection	Initial population	Generation	Optimal iteration	RMSEC	RC	RMSEV	RV	RMSEP	RP
Boiling point	15	100	20	10	3.5551	0.9781	3.5277	0.9732	4.5631	0.9863
		400	43	9	3.2307	0.9820	2.9090	0.9817	4.6179	0.9861
	50	100	120	7	2.4102	0.9901	2.4009	0.9870	4.1999	0.9868
		400	220	8	2.4261	0.9900	2.1553	0.9896	4.2915	0.9855
	200	100	210	8	1.9216	0.9937	2.0636	0.9903	4.4851	0.9842
		400	300	7	2.0428	0.9929	1.8830	0.9920	4.1291	0.9866
	Full spectrum	-	-		1.4898	0.9962	2.8482	0.9923	4.2219	0.9881
Hemoglobin	15	100	62	5	0.3357	0.9805	0.5090	0.9856	0.2878	0.9812
		400	151	5	0.3184	0.9825	0.4700	0.9871	0.2673	0.9837
	50	100	120	5	0.2925	0.9852	0.4222	0.9897	0.2653	0.9840
		400	202	6	0.2868	0.9857	0.4117	0.9901	0.2409	0.9868
	200	100	155	6	0.2495	0.9892	0.3595	0.9923	0.2317	0.9878
		400	260		0.2386	0.9901	0.3253	0.9937	0.2200	0.9891
	Full spectrum	-	-		0.2063	0.9926	0.5673	0.9853	0.2408	0.9868

4. Conclusion

This study proposes and evaluates SLM-GA as multivariate calibration model in near infrared spectroscopic analysis using two experimental dataset. The proposed model has generated improved predicting outcomes for boiling point and hemoglobin prediction. In addition, the proposed model has reduced the number of wavelengths and has higher accuracy than the models that used the full spectrum. Also, SLM-GA outperforms PLS-GA for predicting both datasets and the result indicates that SLM without genetic algorithm has higher result than PLS-GA. This result considered a new investigation which is predetermined wavelengths. The result has shown that predetermined wavelengths provide a considerable result which performs better and similar result than the use of full spectrum when 50 and 200 wavelengths were used where 50 wavelengths were considered a lower number of wavelength compare than 400 and 700 wavelengths for boiling point and hemoglobin full spectrum, respectively. Findings suggest that the proposed SLM-GA is insusceptible to the number of iterations when the SLM was trained with excessive iteration after the optimal iteration number.

Acknowledgment

The authors would like to acknowledge Universiti Tun Hussein Onn Malaysia for providing facilities for this study; and the Southwest Research Institute of San Antonio, Texas, and Eigenvector Research, Inc. (Manson, Washington) for the datasets used in this study

References

[1] Koljonen J, Nordling TE, Alander JT. A review of genetic algorithms in near infrared spectroscopy and chemometrics: past and future. *Journal of Near Infrared Spectroscopy*. 2008; 16(3):189-97.

[2] Bedboudi A, Bouras C, Kimour MT. An Heterogeneous Population-Based Genetic Algorithm for Data Clustering. *Indonesian Journal of Electrical Engineering and Informatics*. 2017; 5(3):275-84.
[3] Khalid S. Performance Evaluation of GA optimized Shunt Active Power Filter for Constant Frequency Aircraft Power System. *Indonesian Journal of Electrical Engineering and Informatics*. 2016; 4(2):112-9.
[4] Umar S, Sridevi G. Analysis of Genetic Algorithm for Effective Power Delivery and with Best Upsurge. *Indonesian Journal of Electrical Engineering and Informatics*. 2017;5(3):264-9.
[5] Arakawa M, Yamashita Y, Funatsu K. Genetic algorithm-based wavelength selection method for spectral calibration. *Journal of Chemometrics*. 2011;25(1):10-9.
[6] Xie H, Zhao J, Wang Q, Sui Y, Wang J, Yang X, Zhang X, Liang C. Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis. *Scientific reports*. 2015; 5:10930.
[7] Feng YZ, Sun DW. Near-infrared hyperspectral imaging in tandem with partial least squares regression and genetic algorithm for non-destructive determination and visualization of Pseudomonas loads in chicken fillets. Talanta. 2013; 109:74-83.
[8] Soares AS, de Lima TW, Soares FA, Coelho CJ, Federson FM, Delbem AC, Van Baalen J. Mutation-based compact genetic algorithm for spectroscopy variable selection in determining protein concentration in wheat grain. *Electronics Letters*. 2014; 50(13): 932-4.
[9] Lavine BK, White CG. Boosting the Performance of Genetic Algorithms for Variable Selection in Partial Least Squares Spectral Calibrations. *Applied spectroscopy*. 2017; 71(9): 2092-2101.
[10] Silalahi DD, Reaño CE, Lansigan FP, Panopio RG, Bantayan NC. Using genetic algorithm neural network on near infrared spectral data for ripeness grading of oil palm (Elaeis guineensis Jacq.) fresh fruit. *Information Processing in Agriculture*. 2016; 3(4): 252-61.
[11] Shan H, Fei Y, Huan Y, Feng G, Fei Q. Quantitative near-infrared spectroscopic analysis of trimethoprim by artificial neural networks combined with modified genetic algorithm. *Chemical Research in Chinese Universities*. 2014; 30(4): 582-6.
[12] Al-kaf HA, Chia KS, Alduais NA. A comparison between single layer and multilayer artificial neural networks in predicting diesel fuel properties using near infrared spectrum. *Petroleum Science and Technology*. 2018; 36(6): 411-8.
[13] Xiaobo Z, Jiewen Z, Povey MJ, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*. 2010; 667(1-2): 14-32.
[14] B. I. Fred McClure, *Chemometrics ShootOut Rules*, International Diffuse Reflectance Conference 2012, 2012.
[15] Mohd Nazrul Effendy Mohd Idrus and Kim Seng Chia. Partial Least Square with Savitzky Golay Derivative in Predicting Blood Hemoglobin Using Near Infrared Spectrum. *MATEC Web Conf.* 2018; 150(01001):1-6.
[16] Babatunde OH, Armstrong L, Leng J, Diepeveen D. Zernike moments and genetic algorithm: Tutorial and application. *British Journal of Mathematics & Computer Science*, 2014; 4(15): 2217-2236.
[17] Yarpiz. (2018). Binary and Real-Coded Genetic Algorithms in MATLAB-Yarpiz. [online] Available at: http://yarpiz.com/23/ypea101-genetic-algorithms [Accessed 20 Feb. 2018]