ICIC Express Letters
Volume **10**, Number **1**, January **2016**

 ISSN 1881-803X
pp. **21–26**

INCORPORATING PRIOR KNOWLEDGE IN SOLVING SYSTEM IDENTIFICATION PROBLEM WITH INSUFFICIENT SAMPLES BASED ON PARETO OPTIMALITY CONCEPT

Mohd Ibrahim Shapiai[1], Zuwairie Ibrahim[2], Asrul Adam[3] and Norrima Mokhtar[3]

[1]Malaysia-Japan International Institute of Technology
Universiti Teknologi Malaysia
Jalan Semarak, Kuala Lumpur 54100, Malaysia
md_ibrahim83@utm.my

[2]Faculty of Electrical and Electronic Engineering
Universiti Malaysia Pahang
Pekan 26600, Pahang, Malaysia
zuwairie@ump.edu.my

[3]Applied Control and Robotics (ACR) Laboratory
Department of Electrical Engineering
Faculty of Engineering
University of Malaya
Kuala Lumpur 50603, Malaysia
asrul.adam@siswa.um.edu.my; norrimamokhtar@um.edu.my

Received June 2015; accepted August 2015

Abstract. *Non-linear modeling based on limited samples is a difficult problem. Incorporating a prior knowledge to this type of problem might offer a promising solution. Various techniques have been proposed to incorporate prior knowledge but depend on one optimal solution which subject to pre-selection of coefficients. Incorporating the knowledge based on Pareto optimality concept offers simple post-selection of solutions. Yet, the proposed Pareto optimality concept may trap to either under-fitting or over-fitting problem based on the obtained Pareto front. The focus of this study is primarily to improve the initialization of the chromosome in order to obtain a reliable Pareto front. One system identification of control engineering problem is used as a problem to be validated. It is shown that the proposed technique is possible to be implemented by capturing the best solution in the obtained Pareto front and relatively improve the accuracy up to 8% performance of the prediction.*
Keywords: Pareto optimality, Prior knowledge, Small samples, System identification

1. **Introduction.** Obtaining adequate data samples is necessary for model generalization especially in a context of the regression problem. However, the data samples collection is costly and time consuming. Recently, the application of learning from small samples has gained increasing attention in many fields, such as in semiconductor biological studies [1], and engine control simulation [2]. There are numerous techniques in machine learning for solving regression problem. However, most of the available techniques mainly focus on solving sufficient training samples problem. Incorporating a prior knowledge is a plausible method in facilitating the data-driven technique to improve the quality of the model.

In general, the existing methods in incorporating prior knowledge rely on one optimal solution which requires the pre-selection of coefficients in formulating the learning function of the regression algorithm [2]. As a result, the existing approaches may simply ignore the possible consequences of the pre-selection of coefficients as limited knowledge is accessible. In view of this limitation, an alternative method to incorporate prior knowledge based on

Pareto optimality has been proposed [3]. Yet, the proposed technique suffers from various problems such as uncertainty of obtained Pareto front and complexity of the problem space.

In this study, the improvement will be based on finding the best setting and reliable Pareto front. The chromosome initialization based on genetic algorithm (GA) multi-objective technique is not randomly initialized but based on regularization function in order to obtain the best setting. Also, obtaining a reliable Pareto is performed by introducing and archive through series of runs before obtaining the Pareto front through non-dominated sorting.

The next section briefly describes the employed techniques in this study. Section 3 explains the details of the proposed technique. The experimental setup, results and discussions are presented in Section 4. Finally, Section 5 is the conclusions.

2. **Employed Techniques.** The weighted kernel regression (WKR) [4] is introduced to solve small sample problems by mapping the input data into the kernel space. Given training samples, $S = \{s_i : X_i,\ y_i | i = 1, 2, \ldots, n\}$, where X_i, $X_i \in \Re^d$ is used to denote the input space (independent variable(s)), y_i, $y_i \in \Re$ is used to denote the output domain (dependent variable), where d refers to the dimensional size of the input space, and n is the number of available training samples. The input mapping is an important element to be used in the proposed technique to transform the linear observed samples to non-linear problems and facilitates the non-linear modeling. In general, WKR required a training stage, i.e., weight estimation before predicting the test sample based on Equation (1)

$$\min f(\alpha) \Leftrightarrow \min \|K(X, X)\,\alpha - y\|^2 \tag{1}$$

where $K(X, X)$ is a square matrix, α is a weight parameter to be estimated and y is the given target output.

Incorporating ridge regression (RR) to WKR was first introduced in [5] to extend the capability of WKR when dealing with noisy samples. The RR is introduced in WKR by adding the L2 regularization term to Equation (1) as given in Equation (2) in order to avoid the singular matrix problem [6]. This is also to ensure a lower variance model by compromising between solving the equation and at the same time keep the α small.

$$f_{reg}(\alpha) = \|K(X, X)\alpha - y\|^2 + \lambda\|\alpha\|^2 \tag{2}$$

where λ is a positive constant value. Differentiating Equation (2) with respect to α gives the closed form solution in estimating the weight parameter as given in Equation (3).

$$\alpha_{ridge} = \left[K(X, X)^T K(X, X) + \lambda I\right]^{-1} K(X, X)^T y \tag{3}$$

where α_{ridge}, $\alpha_{ridge} \in \Re^{n\times 1}$, is the estimated parameters in WKR-RR, and λ is a pre-defined value to control the generalization of the regressed function. In this study, WKR-RR plays an important role in initializing the population of the chromosome in obtaining the Pareto front. Once the weight parameter is obtained, the model is ready to predict any unseen samples (test samples). The testing samples are denoted by $T = \{t_q : X_{test,q}, y_{test,q} | q = 1, 2, \ldots, l\}$, where $X_{test,q}$, $X_{test,q} \in \Re^d$ is used to denote input space, $y_{test,q}$, $y_{test,q} \in \Re$ is output space for testing sets, and l is the number of testing samples.

In general, a multi-objective optimization algorithm (MOEA) consists of several objectives that are conflicting with one another and the aim is to optimize each of them simultaneously. This is the primary feature to be utilized in the proposed technique. There exist various MOEAs in literature. As non-dominated sorting genetic algorithm II (NSGA-II) offers a better spread of solutions, converge better in the obtained Pareto front through a diversity preservation mechanism [7]. Thus, we employed the NSGA-II in the proposed technique. However, the NSGA-II is prone to high uncertainty due to the stochastic nature of the algorithm which may cause unreliable Pareto front for small dataset problem.

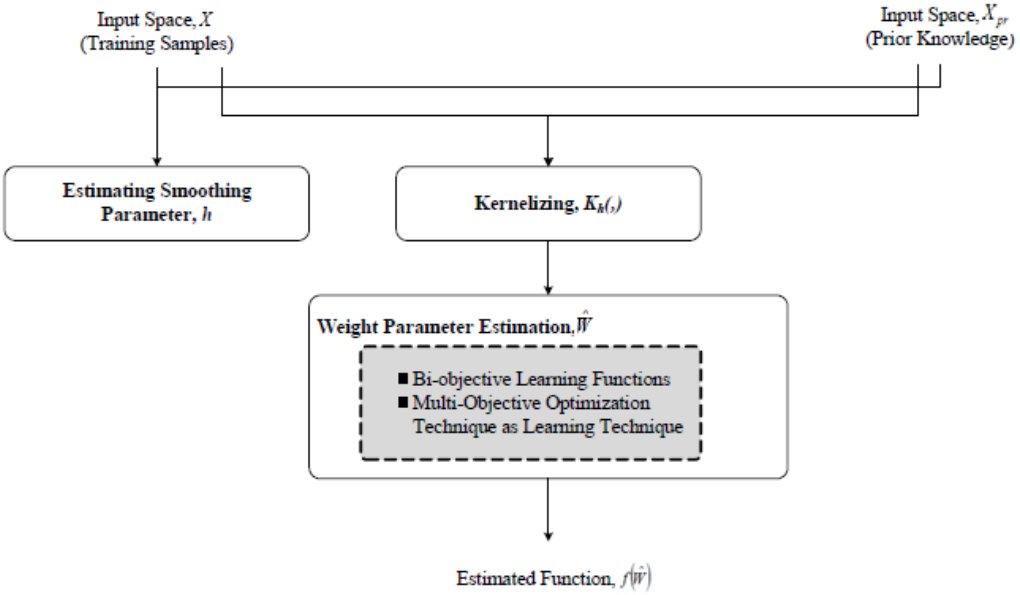

FIGURE 1. Framework of the proposed technique in incorporating prior knowledge

3. **Proposed Technique.** In this section, the framework of the proposed technique will be explained based on Figure 1. Prior to the explanation of the proposed technique, another important notations are introduced. Prior knowledge as an additional information is defined as follows:

$$R = \left\{ r_{i_p} : X_{pr,i_p}, y_{pr,i_p} | i_p = 1, 2, \ldots, n_p \right\} \tag{4}$$

where X_{pr,i_p}, $X_{pr,i_p} \in \Re^d$ refers to the input space and y_{pr,i_p}, $y_{pr,i_p} \in \Re$ refers to the output space. i_p and n_p refer to the ith-index and the number of prior knowledge, respectively.

The proposed technique consists of three main blocks: (1) smoothing parameter estimation, (2) kernelizing and (3) weight parameter estimation. A similar process as in WKR [8] except the proposed technique deals with two conflicting objective functions.

3.1. **Formulation of learning functions.** The formulation of bi-objective learning function is derived based on the available training samples and prior samples. The first and second formulated bi-objective learning functions, f_1 and f_2, are given in Equation (5) and Equation (6)

$$\begin{aligned} f_1(W) = \arg\min_W \Big[& 1/nc_1^{train} \left(K(X_c, X)^{\mathrm{T}} W - Y \right)^2 \\ & + 1/n_p c_2^{train} \left(K(X_c, X_{pr})^{\mathrm{T}} W - Y_{pr} \right)^2 \Big] \end{aligned} \tag{5}$$

$$\begin{aligned} f_2(W) = \arg\min_W \Big[& 1/n_p c_1^{prior} \left(K(X_c, X_{pr})^{\mathrm{T}} W - Y_{pr} \right)^2 \\ & + 1/nc_2^{prior} \left(K(X_c, X)^{\mathrm{T}} W - Y \right)^2 \Big] \end{aligned} \tag{6}$$

where $K(X_c, X)$ maps the available training samples based on the combined samples from $\Re^{n \times d}$ to $\Re^{n_t \times n}$, Y is the output domain value for training samples, $K\left(X_c, X_{pr}\right)$ maps the available prior knowledge based on the combined samples from $\Re^{n_p \times d}$ to $\Re^{n_t \times n_p}$, Y_{pr} is the output domain value for prior knowledge, W, $W \in \Re^{n_t \times 1}$ refers to weight parameters value to be estimated, and c_1^{train} and c_2^{train} are the two coefficients to be pre-defined, where $c_1^{train} + c_2^{train} = 1$. The needs of $1/n$ and $1/n_p$ are to normalize the two terms of the formulated first bi-objective learning function which is important in obtaining the intended solution.

3.2. **Estimation of weight parameters using NSGA-II.** Once the bi-objective learning function is formulated, the weight parameter, W is ready to be estimated. Initially, the NSGA-II parameters, λ value, and number of runs, r have to be defined. The NSGA-II parameters including population size, number of generation, probability of cross-over, p_c, and probability of mutation, p_m are defined to execute the algorithm in obtaining Pareto front. Meanwhile, the determination of λ value is primarily introduced to initialize the population of the chromosome in NSGA-II. The number of runs, r is defined to iteratively execute NSGA-II to reduce the uncertainty in obtaining good and reliable Pareto front.

Once the initialization of the involved parameters is defined, there are two main processes involved in this second phase of the weight parameter estimation block. The first process is introduced to find the best setting of λ value in initializing the chromosome population of the NSGA-II in avoiding the obtained Pareto front to be trapped either in globally or locally Pareto front. Therefore, in this first process, the population initialization is introduced based on WKR as given in Equation (7)

$$W_{init} = \left(K_c^T K_c + \lambda I\right)^{-1} K_c^T Y_c \tag{7}$$

Meanwhile, the second process is introduced to find a reliable and good Pareto front by utilizing the obtained best setting in the previous process through several runs.

4. **Experimental Setup, Results and Discussion.** This section demonstrates the feasibility and effectiveness of the proposed technique to solve system identification problem of control engineering problem. There exist several techniques to perform the system identification which focus on the data-driven approaches such as Artificial Neural Networks (ANNs) [9], and Adaptive Neuro-Fuzzy Inference System (ANFIS) [10].

In this study, PT 326 Process Trainer models is used as a benchmark problem. The device's function is like a hair dryer [10]. The air is flowing in the tube by a centrifugal blower and heated at the inlet. The heated passes through a heater grid before being released into the outlet and measured by a thermocouple. The input, $u(k)$ is the voltage over a mesh of resistor wires to heat incoming air and the output, $v(k)$, is the outlet air temperature. Figure 2 illustrates a block diagram model of the PT 326 Process Trainer.

Firstly, the experimental dataset is obtained from [10,11]. As in [10], the dataset was divided into two sets: (1) training samples ($k = 1$ to 300), S and (2) testing samples ($k = 301$ to 600), T. The input is partitioning into two disjoint sets:

$$input = \begin{cases} V = (v(k-1), v(k-2), v(k-3), v(k-4)) \\ U = (u(k-1), u(k-2), u(k-3), u(k-4), u(k-5), u(k-6)) \end{cases} \tag{8}$$

In this study, input space of the training samples, $X = \{v(k-1), v(k-2), u(k-3)\}$ and the corresponding output space $Y = v(k)$ are selected similarly to [10]. Meanwhile, the simulation data is generated through the Simulink in Matlab by keeping a similar parameter setup as in real experiment. Therefore, the input data to the simulation block

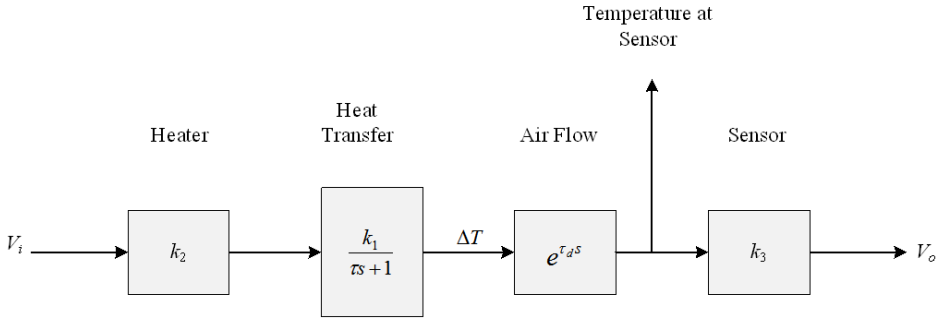

FIGURE 2. Block diagram of the PT326 Process Trainer

TABLE 1. Parameter settings for the conducted experiment of the proposed technique

Parameter	**Values**
WKR Parameter	$h = \max\left(\|X_{k+1}\|^2 - \|X_k\|^2\right)$ where $1 < k < n-1$ and $\|X_{k+1}\|^2 > \|X_k\|^2$
NSGA-II	$c_1^{train} = 1{-}c_2^{train}$, $c_2^{train} = 1e{-}5$, $c_1^{prior} = 1{-}c_2^{prior}$, $c_2^{prior} = 1e{-}5$, $\xi = 0.1$,
Parameters	population size = 100, generation = 100 and iteration, $r = 10$

TABLE 2. RMSE of four different models

Model/Number of Training Samples	**10**	**50**	**100**	**150**	**200**	**250**	**300**
Mixed Model	0.1355	0.1051	0.0913	0.0897	0.0845	0.0780	0.0739
Train Model	0.8977	0.1269	0.0856	0.0727	0.0634	0.0555	0.0505
Proposed Model	0.1341	0.0618	0.0611	0.0512	0.0511	0.0488	**0.0483**
Jang [10]	–	–	–	–	–	–	0.0524

was chosen to be a pseudo random binary signal (PRBS) shifting between 3.5 V and 6.5 V. A similar input selection and output selection as in experimental dataset are employed in determining the prior knowledge.

Seven experiments are conducted with different number of experimental data with fixed number of simulation data. The seven different number of experiment data is assigned as follows, $n = \{10, 50, 100, 150, 200, 250, 300\}$ at one particular number of simulation data, $n_p = 50$. The objective of fixing the number of simulation data is to investigate the importance of the simulation data as the experimental data is increased to maximum samples number. Common parameter's settings for the proposed technique are given in Table 1.

As in [10], the same testing samples are used in validating the proposed model. The quality performance of train model, mixed model to our proposed model is compared. This is to emphasize the need of prior knowledge and how to treat the training samples and prior knowledge in dealing with small sample problem. The train model is developed by using Equation (2) on the training samples only, and a mixed model is developed by Equation (2) on the training samples simply extended with the prior knowledge. The performance of the proposed model is also specifically compared to ANFIS [10] at $n = 300$.

In all experiments, root mean squared error (RMSE) is used as performance index. The same parameter settings as given in Table 1 are employed to perform all models except Jang model [10]. The quality performance for every model is tabulated in Table 2. In general, the available prior knowledge considerably improved the regression quality as it covers a wider region of the input space especially when the experimental data is small. However, the mixed model only recorded very little improvement in terms of RMSE even though number of training samples is increased. The mixed model is simply averaging the available information from the prior knowledge. Lastly, the proposed technique also recorded a slight improvement of the testing accuracy which is up to 8% as compared to [10] at $n = 300$.

5. **Conclusions.** An adequate sample is important besides an appropriate hypothesis for model generalization. Usually the data sampling process is time-consuming and cumbersome. One of the plausible methods to address the problem is by incorporating prior knowledge in facilitating the model generalization. In this study, the proposed technique based on Pareto optimality concept by improving the uncertainty of obtained Pareto front and complexity of problem space offers an improvement to obtain a reliable Pareto front. However, the use of prior knowledge is shown to be inappropriate with larger samples but significantly important for small sample problems. Finally, in future, the selection of best

solution will be investigated and refinement of weight estimation process will be further improved. Also, the generation of artificial samples from the obtained Pareto front will be investigated.

Acknowledgment. This work is financially supported by Fundamental Research Grant Scheme (FRGS), VOTE 4F331 from Ministry of Education, Malaysia.

REFERENCES

[1] R. Andonie, L. Fabry-Asztalos, C. Abdul-Wahid, S. Abdul-Wahid, G. Barker and L. Magill, Fuzzy ARTMAP prediction of biological activities for potential HIV-1 protease inhibitors using a small molecular dataset, *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, 2009.

[2] G. Bloch, F. Lauer, G. Colin and Y. Chamaillard, Support vector regression from simulation data and few experimental samples, *Information Sciences*, vol.178, pp.3813-3827, 2008.

[3] M. I. Shapiai, Z. Ibrahim and M. Khalid, Enhanced weighted kernel regression with prior knowledge using robot manipulator problem as a case study, *Procedia Engineering*, vol.41, pp.82-89, 2012.

[4] M. I. Shapiai, Z. Ibrahim, M. Khalid, L. W. Jau, V. Pavlovic and J. Watada, Function and surface approximation based on enhanced kernel regression for small sample sets, *International Journal of Innovative Computing, Information and Control*, vol.7, no.10, pp.5947-5960, 2011.

[5] M. I. Shapiai, Z. Ibrahim, S. Sudin and M. Khalid, Investigation on different learning techniques for weighted kernel regression in solving small sample problem, *ICIC Express Letters*, vol.6, no.3, pp.705-711, 2012.

[6] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, New York, USA, 2004.

[7] K. Deb, S. Agrawal, A. Pratap and T. Meyarivan, A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II, *Lecture Notes in Computer Science*, vol.1917, pp.849-858, 2000.

[8] M. I. Shapiai, Z. Ibrahim, M. Khalid, L. W. Jau, S. C. Ong and V. Pavlovich, Recipe generation from small samples by weighted kernel regression, *International Conference on Modeling Simulation and Applied Optimization*, Kuala Lumpur, Malaysia, pp.1-4, 2011.

[9] S. Chen, S. Billings and P. Grant, Non-linear system identification using neural networks, *International Journal of Control*, vol.51, pp.1191-1214, 1990.

[10] J. S. R. Jang, Neuro-fuzzy modeling for dynamic system identification, *Proc. of the 1996 Asian Fuzzy Systems Symposium on Soft Computing in Intelligent Systems and Information Processing*, pp.320-325, 1996.

[11] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall PTR, Upper Saddle River, NJ, USA, 1999.