Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 41 (2012) 82 – 89

Engineering Procedia

www.elsevier.com/locate/procedia

International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012)

Enhanced Weighted Kernel Regression with Prior Knowledge Using Robot Manipulator Problem as a Case Study

Mohd Ibrahim Shapiai[a]*, Zuwairie Ibrahim[b], Marzuki Khalid[c]

[a]*Centre of Artificial Intelligent and Robotics (CAIRO), Universiti Teknologi Malaysia, 54100, Kuala Lumpur, Malaysia.*
[b]*Faculty of Electrical & Electronic Engineering,Universiti Malaysia Pahang, Kampus Pekan, 26600 Pekan, Pahang, Malaysia.*

Abstract

Previously, weighted kernel regression (WKR) for solving small samples problem has been reported. In general, WKR has proven to be effective when learning from small samples as compared to artificial neural network with back-propagation (ANNBP) and some other techniques. In order to extend the capability of the technique, we introduce a new approach to improve the WKR by incorporating the prior knowledge. In practice, different forms of prior knowledge may be available and it might avoid the weakness of the training samples limitation. In this study, the incorporation of the prior knowledge will produce a set of solutions by considering the available training samples and prior knowledge in modeling. The process involved in obtaining a set of solutions can be regarded as a bi-objective optimization problem. The proposed technique is derived based on the pareto optimality concept (POC) by using multi-objective optimization technique (MOPT). We only focus the study on the challenges of formulating the two objective functions. We demonstrate the capability of the proposed technique to robot manipulator problem. It is shown that the incorporation of the prior knowledge based on POC can be implemented and relatively improved the regression performance. Some related issues of the proposed technique are also discussed.

Keywords: Small Samples; Weighted Kernel Regression; Pareto Optimality; Prior Knowledge

1. Introduction

Obtaining adequate data samples are necessary for model generalization especially in a context of the regression problem. However, the data samples collection are costly [1] and time consuming [2]. Recently, the application of learning from small samples has gained increasing attention in many fields, such as in semiconductor manufacturing [3], biological studies [4], and engine control simulation [5]. There are numerous techniques in machine learning for regression. However, most of the available techniques mainly focus in solving sufficient training samples problem.

In general, several techniques have been introduced to overcome the limits of learning from small data samples such as finely tune the model parameter [6], pre-data processing [4, 7] and incorporation of prior knowledge [2, 5, 8]. However, there is no universally optimal solution to this problem [4]. Each technique has the capability in solving the problems as compared to the data-driven approaches in a black box modeling technique.

Previously, weighted kernel regression (WKR) has proven to solve small sample with good accuracy for theoretical functions and application in semiconductor problem. The former solution of the WKR is based on the tuning of the model parameter [6]. While the later solution is relying on the pre-data processing technique [7]. Incorporating a prior knowledge

* Corresponding author. Tel.: +6-03-26913710; fax: +6-03-26970815.
E-mail address: ibrahimfke@gmail.com

1877-7058
doi:10.1016/j.proeng.2012.07.146

is a plausible method in facilitating the data-driven approach to improve the quality of the model [2, 5]. In general, the incorporation of the prior knowledge can benefit several engineering problems including robotics [9]. In the mobile robot problems, the prior knowledge may available in a form of action and observation effects of the environment [9], human expertise [10] and items features in the environment [11].

Thus, in this study, we try to incorporate a prior knowledge to the WKR based on the pareto optimality concept (POC). The POC is derived from the multi-objective optimization technique (MOPT). Basically, the main idea of the proposed technique is based on the assumption that any observed target outputs are contaminated by mean-zero additive Gaussian noise with standard deviation. This is the feature that can be utilized in executing the idea of the proposed technique. By nature, the MOPT also offers the trade-off between two or more objective functions. With this, formulating the objective functions of the proposed technique is important in order to make use of the POC.

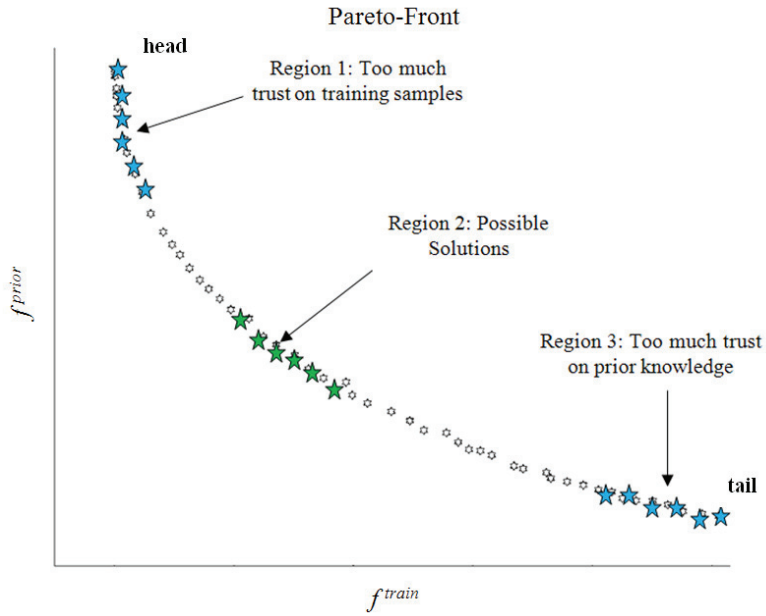


Fig. 1. Solutions lying on the Pareto-Front

Based on the proposed idea, the possible solutions that lie on the Pareto-Front can be divided into three regions as shown in Fig 1. The first two regions lie close to the two extreme points i.e. head and tail and the third region lie in the middle of the Pareto-Front. The solutions on the head and tail of Pareto-Front region are biased to training samples and prior knowledge respectively. Ideally, reliable solution may exist in the middle of the Pareto-Front.

In this paper, the WKR is used to map all the training samples and prior knowledge into kernel space [6-7]. The two formulated objective functions in kernel space are then solved using MOPT in order to generate the Pareto-Front. Prior to the generation of the Pareto-Front, it is necessary to initialize the population of the chromosomes in MOPT. In this study, the population initialization is performed using a WKR - Ridge Regression (RR) (WKR-RR) technique [12]. In summary, there are three main techniques employed in implementing the proposed technique.

This paper is organized as follows. A brief review of all employed techniques is given in section 2. The proposed technique is formulated in section 3. The setup experiments and results of the proposed technique are discussed in section 4. Finally, the conclusion is discussed in section 5.

2. Employed techniques

The WKR [6-7] is introduced to solve small sample problems by mapping the input data into the kernel space. The input mapping is 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. The WKR is a modified Nadaraya-Watson kernel regression (NWKR) [13] by expressing the observed samples in a square kernel matrix. As compared to NWKR, WKR required a training stage, i.e. weight estimation before predicting the test sample based on the Equation (1)

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

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 [12] 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 [14]. This is also to ensure a lower variance model by compromising between solving the equation and at the same time keep the α small.

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

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 $\alpha_{ridge}, \alpha_{ridge} \in \Re^{n\times 1}$, is the estimated parameters in WKR-RR, λ is a predefined 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 MOPT.

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 based on POC. There exist various MOEAs in literatures in the last two decades. As non-dominated sorting genetic algorithm II (NSGA-II) [15] offers a better spread of solutions, converge better in the obtained Pareto-Front through a diversity preservation mechanism [16] . Thus, we employed the NSGA-II in the proposed technique.

3. Proposed technique

Basically, the proposed technique consists of three main blocks is shown as in Figure 2: (1) Training Samples, (2) Prior Knowledge and (3) MOPT block.

3.1. Training samples block

Initially, the observed training samples, $\left\{ X_i^t, y_i^t \right\}$ for $i = 1,\ldots,n_t$ where $X_i^t \in \Re^d$ and $y_i^t \in \Re^{n_t \times 1}$ and prior knowledge, $\left\{ X_i^p, y_i^p \right\}$ for $i = 1,\ldots,n_p$ where $X_i^p \in \Re^d$ and $y_i^p \in \Re^{n_p \times 1}$ are simply concatenated which are defined as combined samples, $\left\{ X_i^c, y_i^c \right\}$ for $i = 1,\ldots,n$ where n is a summation of n_t and n_p, $X_i^c \in \Re^d$ and $y_i^c \in \Re^{n\times 1}$. The concatenated samples are required to ensure the proposed technique has enough free parameters (estimated weight). The inclusion of the prior knowledge facilitates the modeling especially when the available prior samples cover a wider region of the input space [8].

The concatenated samples are then mapped into kernel space before formulating the first objective function for the training samples as given in Equation (4)

$$f_1(W) = \min_W \left[\frac{1}{n_t} c_1^{train} \left(K\left(X_i^c, X_i^t\right)^{\mathrm{T}} W - y_i^t \right)^2 + \frac{1}{n_p} c_2^{train} \left(K\left(X_i^c, X_i^p\right)^{\mathrm{T}} W - y_i^p \right)^2 \right] \tag{4}$$

where $f_1(W)$ is the first objective function that is indexed from 1,2,…,*k*,…,*m* and m is the number of generated solutions that are lying on the Pareto-Front. $K\left(X_i^c, X_i^t\right)$ maps $\Re^{n\times n} \times \Re^{n\times n_t}$ into $\Re^{n\times n_t}$, $K\left(X_i^c, X_i^p\right)$ maps $\Re^{n\times n} \times \Re^{n\times n_p}$ into $\Re^{n\times n_p}$, W, $W \in \Re^{n\times 1}$, which are the sharing decision variables to be estimated (free parameters). In the multi-objective optimization block, c_1^{train} and c_2^{train} are the two coefficients to be pre-defined, $c_1^{train} + c_2^{train} = 1$ and c_1^{train} must be set significantly larger than c_1^{train}, $c_1^{train} >> c_2^{train}$.

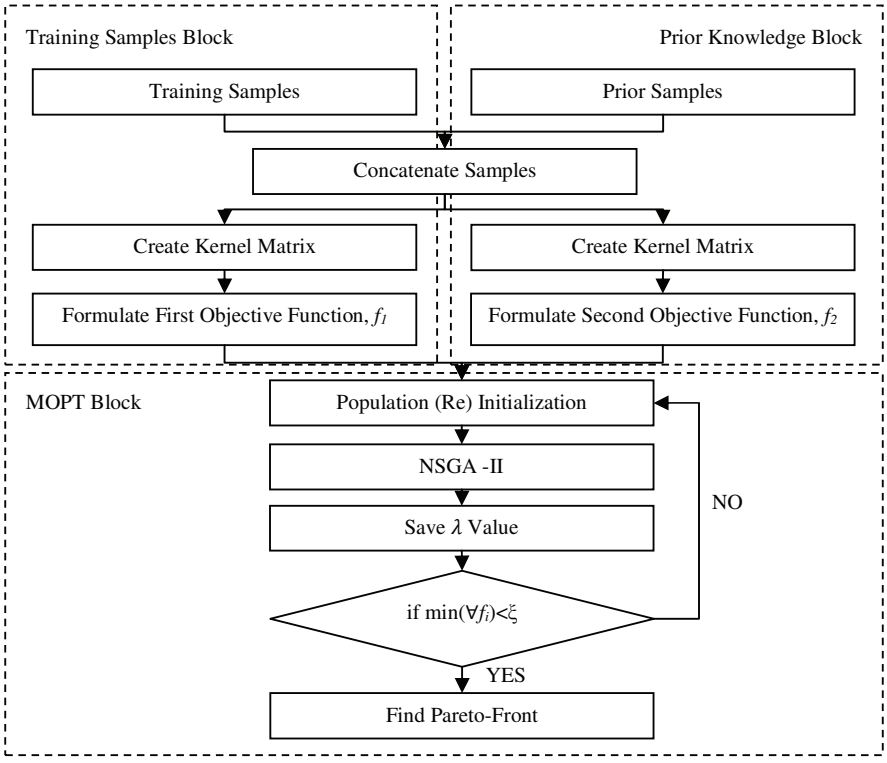


Fig 2. Prior knowledge block

3.2. Prior knowledge block

As in the training samples block, the second objective can be formulated based on the mapping combined samples. The formulated objective function is given by Equation (5) as follows:

$$f_2(W) = \min_W \left[\frac{1}{n_p} c_1^{prior} \left(K\left(X_i^c, X_i^p\right)^{\mathrm{T}} W - y_i^p \right)^2 + \frac{1}{n_t} c_2^{prior} \left(K\left(X_i^c, X_i^t\right)^{\mathrm{T}} W - y_i^t \right)^2 \right] \tag{5}$$

where $f_2(W)$ is the second objective function that are indexed from 1,2,…,*k*,…,*m* and *m* is the number of generated solutions that are lying on the Pareto-Front, $K\left(X_i^c, X_i^p\right)$ maps $\Re^{n\times n} \times \Re^{n\times n_p}$ into $\Re^{n\times n_p}$, $K\left(X_i^c, X_i^t\right)$ maps $\Re^{n\times n} \times \Re^{n\times n_t}$ into $\Re^{n\times n_t}$. $W, W \in \Re^{n\times 1}$, is the sharing decision variable to be estimated (free parameters) in the multi-objective optimization block. c_1^{prior} and c_2^{prior} are the two coefficients to be pre-defined, $c_1^{prior} + c_2^{prior} = 1$ and c_1^{prior} must be set significantly larger than c_1^{prior}, $c_1^{prior} >> c_2^{prior}$.

3.3. Multi-objective optimization block

In this sub-section, the proposed technique will be established using the two previous formulated objective functions. Since the two formulated objective functions are convex functions, there exist many algorithms which can handle multi-objective problems well [17] and their convexity seems to cause the least amount of difficulty in solving problems [15].

Firstly, the population has to be initialized in order to avoid under-fitting or over-fitting of the whole set of solutions in Pareto-Front. The initialization is carried out using the WKR-RR as given in Equation (6). The matrix size of the initial population is based on the number of population and dimension of the combined samples as given in Equation (6)

$$Population\ Init\ =\ \alpha_{ridge} 1^T + \varepsilon \tag{6}$$

where $\alpha_{ridge}, \alpha_{ridge} \in \Re^{n\times 1}$, is the estimated parameters using WKR-RR , $1,1 \in \Re^{n\times 1}$, is column vector of value 1 and ε is the Gaussian error of zero mean and 0.1 standard deviation, $N_{pop}(0,0.1)$.

The initial value of λ in Equation (3) is set to 1, every time the multi-objective algorithm fails to fulfill the error condition, ξ, as given in Equation (7), the λ is decreased by 0.1. As the λ value becomes smaller, the found solution tends to over-fitting yet if the initial value of λ is too large, the founded solution may be trapped into an under-fitting problem. Once the λ is found, the corresponding α_{ridge}, will be used to generate m solutions that lying on the Pareto-Front.

$$\min\left(f_1(W)\right) \wedge \min\left(f_2(W)\right) \leq \xi \tag{7}$$

4. Numerical experiments and results

The selection of the robot manipulator problem as a case study is mainly to exhibit the capability of the proposed technique when incorporating prior knowledge [2]. The robot manipulator problem is generated by using Equation (8).

$$y_i = a_1 \sin\left(\theta_i^{(1)}\right) + a_2 \sin\left(\theta_i^{(1)} + \theta_i^{(2)}\right) + \varepsilon_i \tag{8}$$

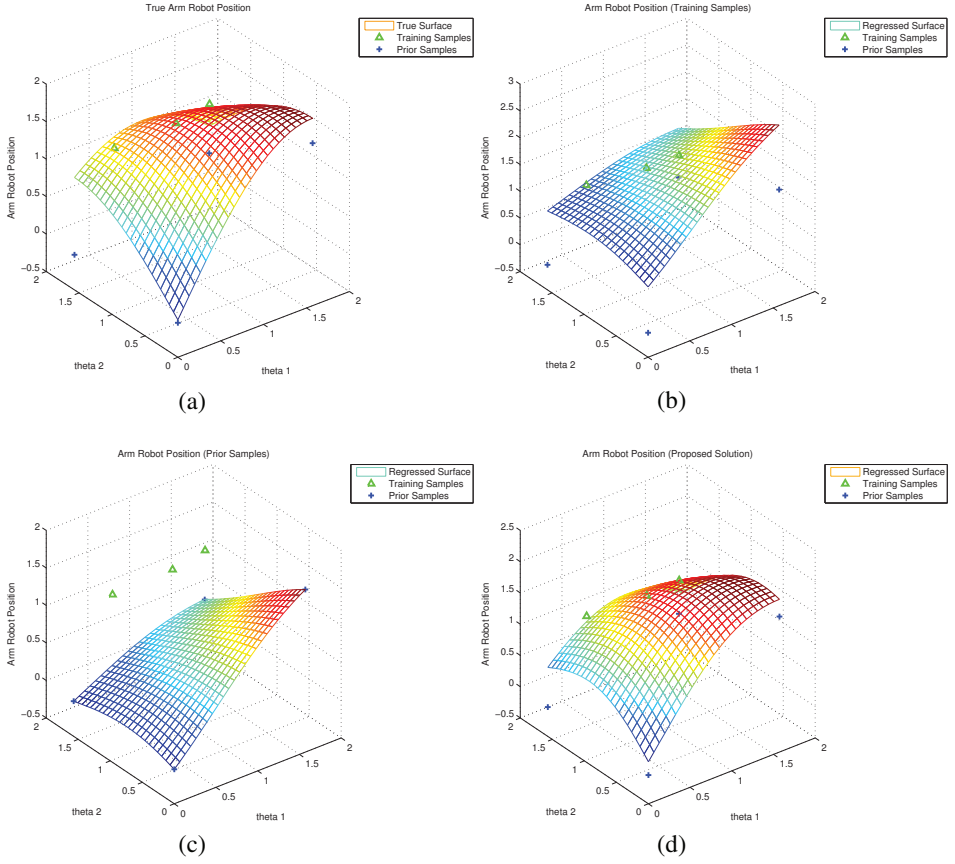


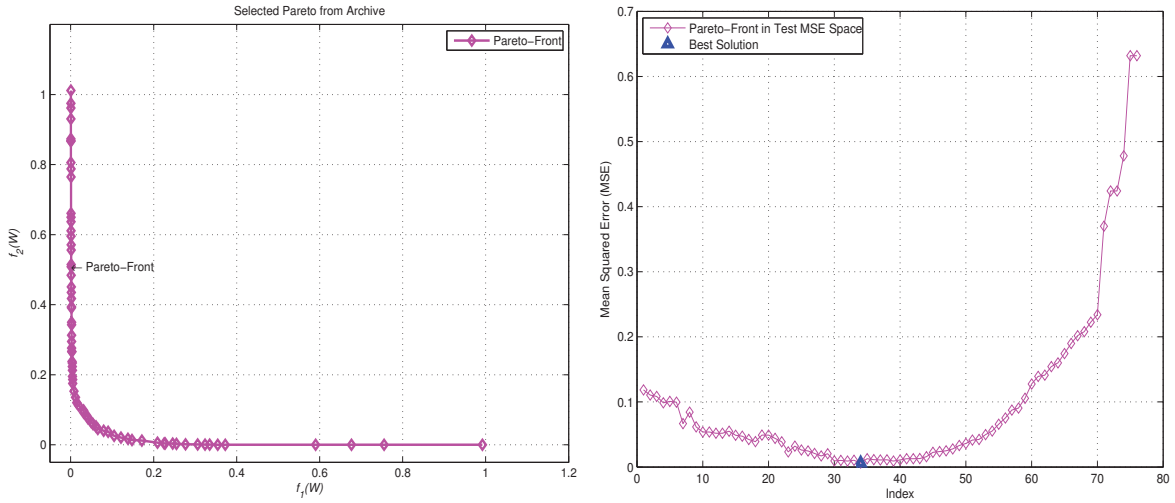
Fig. 3. Experiment results of case 3, (a) the true surface, (b) too much trust on training samples, (c) too much trust on prior samples and (d) the proposed technique based on the best solution.

where y_i is assumed to be corrupted by Gaussian noise with mean zero and standard deviation 0.1~ $N_t(0,0.1)$ for training samples and three different noise setting for prior knowledge which is $N_p(0,0.1)$, $N_p(0,0.3)$ and $N_p(0,0.5)$, $\theta_i^{(1)}$ and $\theta_i^{(2)}$ are drawn uniformly from $[0,\pi/2]\times[0,\pi/2]$ for the training samples and the prior samples which are randomly generated in the edge of the surface as in [2]. We also assume that the generated prior knowledge is a relevant dataset. The parameters a_1 and a_2 are equally set to one. In summary, three different cases are investigated in this study such as follow (1) Case 1: $N_t(0,0.1)$ and $N_p(0,0.1)$, (2) Case 2: $N_t(0,0.1)$ and $N_p(0,0.3)$ and (3) Case 3: $N_t(0,0.1)$ and $N_p(0,0.5)$.

In this experiment, we limit the number of training samples and prior samples into three and four samples, respectively in order to emphasize the effectiveness of the proposed technique when dealing with small samples. The test samples are generated over a grid of 625 points in the set of $\{[0,\pi/2]\times[0,\pi/2]\}$. The distribution of the training and prior samples are shown in Figure 3(a). Initially, the parameter settings of the proposed technique are predefined in Table 1.

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

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



(a) (b)

Fig. 4. Generated solutions (a) Pareto-Front, (a) Test MSE Space derived from Pareto-Front

In all experiments, the performance criterion test MSE, in Equation (9) is used to evaluate the generalization performance of the proposed technique as given in the equation below:

$$MSE=\frac{1}{S}\left(f_{true}-f_{predict}(W)\right)^2 \tag{9}$$

where S is the number of the test data, f_{true} is the true value of the tested function and $f_{predict}(W)$ is the predicted value. In all cases, WKR-RR in Equation (3) is used to regress the function without prior knowledge. The lowest test MSE which corresponds to the chosen λ value is reported in this study. Also, the result from the proposed technique is simply chosen

from the best solution, i.e. the lowest test MSE value from the test MSE space. An approach to find the best solution is not implemented in this study as we leave the task for future work.

A set of solutions that lying on the Pareto-Front is shown in Figure 4(a). We provided the test MSE space that derives from the Pareto-Front by translating every single solution from the Pareto-Front into MSE value with respect to the test samples (unseen samples) by using Equation (9). The test MSE space corresponds to the Pareto-Front and the best solution is shown as in Figure 4(b).

The biased regressed function on training samples which is the founded solution in the extreme point (head), prior samples (tail) and regressed function of the proposed technique are shown in Figure 3(b), 3(c) and 3(d) respectively. The experiment is repeated over ten runs and the measured performances, test MSE are tabulated in Table 2. As noted previously, all the recorded solutions of the proposed technique are based on the best selection from the test MSE space. It is proven that the proposed technique is able to capture lower test MSE region from the Pareto-Front. The proposed technique offers the regularized estimated free parameters as the best solution from the Pareto-Front is not trapped to the under-fitting or over-fitting problem as can be seen from Figure 3(d) and the recorded test MSE in Table 2. Also, the proposed technique is capable to resolve the complexity in regressing the surface with sufficient parameters. The available prior knowledge relatively improved the regression quality as it covers a wider region of the input space.

Table 2. The mean and standard deviation of the test MSE for the robot manipulator function

Problem		MSE
Case 1	Proposed Technique (Best Solution)	0.0097 ± 0.0005
	Without Prior knowledge	0.1900 ± 0.0442
Case 2	Proposed Technique (Best Solution)	0.0153 ± 0.0067
	Without Prior knowledge	0.1896 ± 0.0159
Case 3	Proposed Technique (Best Solution)	0.0192 ± 0.0093
	Without Prior knowledge	0.1959 ± 0.0633

5. Conclusions

An adequate dataset size is important besides an appropriate hypothesis for model generalization. In real problems, usually the data sampling process is time-consuming and cumbersome. One of the plausible methods is incorporating prior knowledge. In summary, incorporating prior knowledge to the WKR raises several issues that have to be foremost considered which are: (1) number of free parameters in formulating the two objective functions and (2) population initialization in MOPT. The first issue is mainly related to the capability of the proposed technique in modeling a complex regression function with sufficient free parameters. Finally, a proper selection in initializing the population is important in order to avoid the problem of under-fitting and over-fitting when estimating weight parameters (free parameters). The two issues are appropriately solved by the proposed technique through a series of experiments, which results in a relatively lower test MSE. In future, the selection of best solution based on the preference information will be investigated and applying the proposed technique in any applications that offers prior knowledge.

Acknowledgements

This work is financially supported by GUP Research Funds (Q.J130000.7109.02H20), Universiti Teknologi Malaysia (UTM) and Ministry of Higher Education Malaysia (MOHE)

References

[1] C. Huang and C. Moraga, "A diffusion-neural-network for learning from small samples," International Journal of Approximate Reasoning, vol. 35, pp. 137-161, 2004.
[2] J. Yuan, et al., "Incorporating prior model into Gaussian processes regression for WEDM process modeling," Expert Systems with Applications, vol. 36, pp. 8084-8092, 2009.
[3] W. Lee and S. Ong, "Learning from small data sets to improve assembly semiconductor manufacturing processes," in 2nd ICCAE 2010, Singapore, 2010, pp. 50-54.
[4] R. Andonie, et al., "Fuzzy ARTMAP prediction of biological activities for potential HIV-1 protease inhibitors using a small molecular dataset," IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2009.
[5] G. Bloch, et al., "Support vector regression from simulation data and few experimental samples," Information Sciences, vol. 178, pp. 3813-3827, 2008.
[6] M. I. Shapiai, et al., "Function and Surface Approximation Based on Enhanced Kernel Regression for Small Sample Sets," International Journal Of Innovative Computing, Information And Control, vol. 7, pp. 5947-5960, 2011.

[7] M. I. Shapiai, et al., "Solving Small Sample Recipe Generation Problem with Hybrid WKRCF-PSO," International Journal on New Computer Architectures and Their Applications, vol. 1, pp. 833 - 843, 2011.
[8] F. Lauer and G. Bloch, "Incorporating prior knowledge in support vector regression," Machine Learning, vol. 70, pp. 89-118, 2008.
[9] W. D. Smart and L. Pack Kaelbling, "Effective reinforcement learning for mobile robots," 2002, pp. 3404-3410 vol. 4.
[10] D. L. Moreno, et al., "Using prior knowledge to improve reinforcement learning in mobile robotics," Proc. Towards Autonomous Robotics Systems. Univ. of Essex, UK, 2004.
[11] Y. Liu, et al., "Using EM to learn 3D models of indoor environments with mobile robots," 2001, pp. 329-336.
[12] M. I. Shapiai, et al., "Investigation on Different Learning Techniques for Weighted Kernel Regression in Solving Small Sample Problem " ICIC Express Letters, An International Journal of Research and Surveys, vol. 6, pp. 705 - 711, March 2012 2012.
[13] È. Nadaraya, "On estimating regression," Teoriya Veroyatnostei i ee Primeneniya, vol. 9, pp. 157-159, 1964.
[14] S. P. Boyd and L. Vandenberghe, Convex optimization. New York, NY, USA: Cambridge Univ Press, 2004.
[15] K. Deb, et al., "A fast and elitist multiobjective genetic algorithm: NSGA-II," Evolutionary Computation, IEEE Transactions on, vol. 6, pp. 182-197, 2002.
[16] K. Deb, et al., "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II," 2000, pp. 849-858.
[17] K. Deb, "Multi-objective genetic algorithms: Problem difficulties and construction of test problems," Evolutionary computation, vol. 7, pp. 205-230, 1999.