# WEIGHTED KERNEL REGRESSION IN SOLVING SMALL SAMPLE PROBLEMS FOR REGRESSION FRAMEWORK AND ITS VARIANTS

MOHD IBRAHIM BIN SHAPIAI @ ABD RAZAK

UNIVERSITI TEKNOLOGI MALAYSIA

# WEIGHTED KERNEL REGRESSION IN SOLVING SMALL SAMPLE PROBLEMS FOR REGRESSION FRAMEWORK AND ITS VARIANTS

MOHD IBRAHIM BIN SHAPIAI @ ABD RAZAK

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy in (Electrical Engineering)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JANUARY 2014

*To my beloved wife and my little daughter.*
*To my parents and siblings.*

## ACKNOWLEDGEMENT

First of all, thanks to ALLAH SWT for granting me the strength and show me the path to complete this research. Also, praise to Prophet Muhammad S.A.W.

Secondly, I am expressing my sincere thanks to Professor Datin Dr. Rubiyah Yusof, the late Professor Datuk Dr. Marzuki and special thanks to Associate Professor Dr. Zuwairie Ibrahim. I deeply appreciate their precious time, guidance, financial support and encouragement in completing this research. I am also not forgetting their thoughts and care especially in reviewing this thesis.

I wish to express my sincere appreciation to Professor Junzo Watada from Waseda University for his guide especially in formulating the ideas during my visit at his lab in April 2010. I would like to express my special gratitude and thanks to Professor Vladimir Pavlovich from Rutgers University and Dr. Lee Wen Jau and his group from Intel Malaysia for their discussion and suggestions. Also, I am highly indebted to Intel Tech Co., Ltd. for their financial support.

My great appreciation is also to all friends in CAIRO especially Rahairi, Amir, Fadhli, Peyo and her wife (Kak Yanti), and many more for their supports and advices.

Furthermore, I would like to acknowledge with much thanks to Universiti Teknologi Malaysia (UTM) and Malaysian government for a study leave and financial support through SLAI-JPA scholarship scheme. Also, not forget to KPT (Scholar Division) for granting a travel grant to present a paper at the EMS2011 conference in Madrid, Spain.

My deepest appreciation is to my wife, Farah Rahim for her support and doa and to my daughter Irdina Zahiah. Last but not least, many thanks to my parents and family members for their support and help.

*Mohd Ibrahim, CAIRO*

# ABSTRACT

In real engineering problems, regression plays an important role as a tool to approximate an unknown target function. One of the primary regression problems is the insufficient number of training samples during regression of a model to find the relationship between input and output of samples. These samples do not provide enough information which leads to difficulty in the regression. Current techniques focusing on the selection of appropriate model parameters use several data-driven techniques and none of these techniques offer generalized and better solutions to the problems. Incorporation of prior knowledge is a plausible technique to improve the quality of the regression for data-driven techniques but this technique is based on pre-selection of coefficients in solving the problem, thus ignoring the consequence of the selection. To improve regression quality for small training samples problems, this thesis explored a new technique known as Weighted Kernel Regression (WKR) based on the Nadaraya-Watson Kernel Regression (NWKR) when solving small training samples problems. Besides WKR, the study introduced the incorporation of prior knowledge based on Pareto-optimality concept that utilizes the genetic algorithm-based multi-objective optimization technique to compute the Pareto optimal solutions. Instead of relying on pre-selection of coefficients to incorporate the knowledge, the proposed technique uses post-selection of solutions as it is less subjective, especially when used for solving small sample problems. The experimental results of the proposed technique using several benchmark problems were evaluated and compared with existing techniques. The findings showed that the proposed technique not only can offer better regression accuracy but also flexibility in generalizing the solution, especially when the available training samples are insufficient.

# ABSTRAK

Dalam masalah kejuruteraan, regresi memainkan peranan yang penting sebagai alat untuk menentukan fungsi sasaran yang tidak diketahui. Salah satu masalah utama regresi adalah bilangan sampel yang tidak mencukupi untuk mencari hubungan antara masukan dan keluaran sampel. Sampel ini tidak memberi maklumat yang mencukupi dan mengakibatkan kesukaran dalam regresi. Kebanyakan teknik semasa menggunakan teknik bersandarkan data dengan hanya memberi tumpuan kepada pemilihan parameter model, namun tiada teknik yang mampu menawarkan penyelesaian umum dan memberi keputusan yang lebih baik kepada masalah yang dikaji. Penggabungan pengetahuan awal adalah teknik yang mampu meningkatkan kualiti regresi tetapi teknik-teknik ini hanya berdasarkan kepada pra pemilihan pekali yang mengabaikan kesan pemilihan pekalinya. Untuk meningkatkan kualiti regresi bagi masalah sampel yang kecil, tesis ini meneroka teknik baru yang dikenali sebagai *Weighted Kernel Regression* (WKR) berdasarkan *Nadaraya-Watson Kernel Regression* (NWKR) dalam menyelesaikan masalah sampel yang kecil. Selain WKR, kajian ini juga memperkenalkan satu teknik baru penggabungan pengetahuan awal yang berasaskan konsep *Pareto-optimality* dengan menggunakan algoritma genetik berasaskan teknik *multi-objective optimization* untuk mengira penyelesaian Pareto yang optimum. Dari bergantung kepada pra pemilihan pekali dalam penggabungan pengetahuan awal, teknik yang diperkenalkan menggunakan pasca pemilihan penyelesaian yang lebih mudah dan sesuai bagi masalah sampel yang kecil. Keputusan eksperimen dari teknik yang dicadangkan dinilai dan dibandingkan dengan menggunakan beberapa masalah penanda aras kepada teknik-teknik yang sedia ada. Hasil kajian menunjukkan bahawa teknik yang dicadangkan bukan hanya boleh menawarkan ketepatan regresi yang lebih baik tetapi juga kefleksibelan dalam menawarkan penyelesaian umum, terutamanya apabila sampel yang ada tidak mencukupi.

# TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| ANN | – | Artificial Neural Network |
| DNN | – | Diffusion Neural Network |
| F-LOOCV | – | Frame based Leave-One-Out Cross-Validation |
| GA | – | Genetic Algorithm |
| GP | – | Gaussian Process |
| MSE | – | Mean Squared Error |
| NA | – | Not Available |
| NSGA-II | – | Non-dominated sorting genetic algorithm II |
| NWKR | – | Nadaraya-Watson Kernel Regression |
| RR | – | Ridge Regression |
| RMSE | – | Root Mean Squared Error |
| SSE | – | Sum of Squared Error |
| SVR | – | Support Vector regression |
| VC | – | Vapnik-Chervonik |
| WEDM | – | Wire-cut Electrical Discharge Machining |
| WKR | – | Weighted Kernel Regression |

## LIST OF SYMBOLS

| | | |
|---|---|---|
| $a_i$ | – | Coefficient for function of arm robot problem |
| $\alpha$ | – | Parameter of parametric regression |
| $b$ | – | Kernel bound |
| $\beta_i$ | – | Spread factor to create new offspring based on crossover operation |
| $c_{i_t}$ | – | $i_t$-th number of combined samples |
| $c_1^{train}$ | – | Pre-defined coefficient for first term of formulated first learning function in multi-objective environment |
| $c_2^{train}$ | – | Pre-defined coefficient for second term of formulated first learning function in multi-objective environment |
| $c_1^{prior}$ | – | Pre-defined coefficient for first term of formulated second learning function in multi-objective environment |
| $c_2^{prior}$ | – | Pre-defined coefficient for second term of formulated second learning function in multi-objective environment |
| $C$ | – | Collection of combined samples |
| $C_i$ | – | Chromosome of population in genetic algorithm |
| $C_i^{(L)}$ | – | The minimum fitness value of any particular chromosome |
| $C_i^{(U)}$ | – | The maximum fitness value of any particular chromosome |
| $d$ | – | Dimensionality size |
| $d_i$ | – | A local crowding distance in particular non-dominated front |
| $\varepsilon$ | – | Gaussian noise |
| $\varepsilon_{init}$ | – | Initialize noise for population initialization |
| $\varepsilon_{pr}$ | – | Gaussian noise for prior knowledge |
| $\xi$ | – | Threshold value |
| $\hat{\xi}$ | – | Estimated threshold value |
| $E(Y\|X)$ | – | Conditional expectation of $Y$ given $X$ |
| $f$ | – | Target function |
| $f_m^{\left(I_j^m\right)}$ | – | The desired solution of the multi-objective problem in objective space |

| | | |
|---|---|---|
| $f_m^{\left(I_{j+1}^m\right)}$ | – | The adjacent solution of the multi-objective problem in objective space with higher value |
| $f_m^{\left(I_{j-1}^m\right)}$ | – | The adjacent solution of the multi-objective problem in objective space with lower value |
| $f_m^{max}$ | – | The maximum obtained value of $m$-th objective function |
| $f_m^{min}$ | – | The minimum obtained value of $m$-th objective function |
| $f_{L_1R_1}$ | – | Learning function with $L_1$ norm error term and $L_1$ norm regularization term |
| $f_{L_2R_1}$ | – | Learning function with $L_2$ norm error term and $L_1$ norm regularization term |
| $f_{L_1R_2}$ | – | Learning function with $L_1$ norm error term and $L_2$ norm regularization term |
| $f_{L_2R_2}$ | – | Learning function with $L_2$ norm error term and $L_2$ norm regularization term |
| $f_{train}(x)$ | – | Target function based on given training samples |
| $f_{prior}(x)$ | – | Target function based on given prior knowledge |
| $\hat{f}$ | – | Most possible function |
| $\hat{f}\left(X_{test,q}\right)$ | – | Estimated of $q$-th input space for testing samples |
| $f(X)$ | – | Function to be estimated |
| $f(\hat{W})$ | – | Estimated function using WKR |
| $f_1(W)$ | – | Formulated first learning function in multi-objective environment |
| $f_2(W)$ | – | Formulated second learning function in multi-objective environment |
| $f_p^{i_{gen}}\left(W\right)$ | – | Threshold function which is based on mean value of half of the updated population of the sorted first learning function |
| $f_{pop/2}^{i_{gen}}\left(W\right)$ | – | Threshold function which is based on middle solution of the sorted first learning function |
| $f_1^{i_{gen}}\left(W\right)$ | – | Threshold function which is based on the minimum solution of the sorted first learning function |
| $f(\alpha, X)$ | – | Parametric function to be estimated |
| $F$ | – | Hypothesis class |
| $g_j\left(W\right)$ | – | Variable of inequality constraint |
| $g(X)$ | – | Non-parametric function to be estimated |
| $gen$ | – | Number of generation for genetic algorithm |
| $h$ | – | Smoothing parameter |
| $h_k\left(W\right)$ | – | Variable of equality constraint |

| | | |
|---|---|---|
| $i$ | – | Index number of training samples |
| $I$ | – | Interval of kernel bound |
| $I_j$ | – | $j$-th solution in the sorted list of the crowding distance algorithm |
| $I_{j-1}$ | – | The adjacent neighbour index of $I_j$ with lower value |
| $I_{j+1}$ | – | The adjacent neighbour index of $I_j$ with higher value |
| $k_{ij}$ | – | Element of kernel matrix for training sample |
| $k_{iq}^{test}$ | – | Element of kernel matrix for testing sample |
| $k_a^n$ | – | $n$-th translation kernel space of the given arbitrary space |
| $K$ | – | Kernel matrix for training sample |
| $K_c$ | – | Kernel space for combined samples with combined samples as reference |
| $K_{tr}$ | – | Kernel space for training samples with combined samples as reference |
| $K_{pr}$ | – | Kernel space for prior knowledge with combined samples as reference |
| $K^{test}$ | – | Kernel matrix for testing sample |
| $K^a$ | – | Kernel space of the given arbitrary space |
| $K_{-i}$ | – | Kernel matrix of input space of $i$-th left-out training set |
| $K_h(.)$ | – | Mapping process from input space to kernel space |
| $K(,)$ | – | Kernel function |
| $l$ | – | Number of testing samples |
| $L_1$ | – | $L_1$ norm |
| $L_2$ | – | $L_2$ norm |
| $f_m$ | – | $m$-th number of objective functions |
| $m_s$ | – | Mass percentage |
| $m_s\,(k+1)$ | – | More mass percentage |
| $m_s\,(k)$ | – | Less mass percentage |
| $M$ | – | Total number of formulated objective functions |
| $\eta_c$ | – | Probability of creating near-parent or distant parent solutions in simulated binary crossover |
| $\eta_m$ | – | Probability of creating near-parent or distant parent solutions in polynomial mutation |
| $n$ | – | Number of training samples |
| $n_p$ | – | Number of prior knowledge |
| $n_t$ | – | Number of combined samples |

| | | |
|---|---|---|
| $N \sim (0, 0.1)$ | – | Gaussian noise with zero mean value and 0.1 standard deviation |
| $N \sim (0, 0.3)$ | – | Gaussian noise with zero mean value and 0.3 standard deviation |
| $N \sim (0, 0.5)$ | – | Gaussian noise with zero mean value and 0.5 standard deviation |
| $p$ | – | $p$-th number of observed dimensionality size |
| $pop$ | – | Number of population in genetic algorithm |
| $p_c$ | – | Probability of cross-over |
| $p_m$ | – | Probability of mutation |
| $q$ | – | Index number testing samples |
| $r$ | – | Number of runs in multi-objective environment |
| $r_i$ | – | Non-dominated fronts ranking for crowding distance algorithm |
| $R$ | – | Collection of prior knowledge |
| $r_{i_p}$ | – | $i_p$-th number of prior knowledge |
| $\Re$ | – | Real space |
| $\delta_i$ | – | Spread factor to create new offspring based on mutation operation |
| $S$ | – | Collection of training samples |
| $s_i$ | – | $i$-th number of training samples |
| $T$ | – | Collection of testing samples |
| $t_q$ | – | $q$-th number of testing samples |
| $\mu_i$ | – | Coefficients parameter |
| $u$ | – | Component of independent variable |
| $U$ | – | Input variables (independent variables) |
| $v$ | – | Output variable (dependent variable) |
| $W$ | – | Weight parameters |
| $W_{init}$ | – | Initialized weight parameters |
| $W_i$ | – | $i$-th number of weight parameters |
| $W_i^{(L)}$ | – | Lower limit value of the estimated weight parameters |
| $W_i^{(U)}$ | – | Upper limit value of the estimated weight parameters |
| $W_{-i}$ | – | Estimated weight parameters of $i$-th left-out training set |
| $W^r$ | – | Current update value of weight parameters |
| $W^{r+1}$ | – | Next update value of weight parameters |
| $\hat{W}$ | – | Estimated weight parameters |
| $x_a^d$ | – | $d$-th element of arbitrary input space for training samples |
| $X$ | – | Input space of training samples |
| $X_a$ | – | Arbitrary input space for training samples |

| | | |
|---|---|---|
| $X_c$ | – | Input space of combined samples |
| $X_i$ | – | $i$-th number of input space for training samples |
| $X_{pr}$ | – | Input space of prior knowledge |
| $X_i^p$ | – | $p$-th dimension of $i$-th number of input space for training samples |
| $X_{c,i_t}$ | – | $i_t$-th number of input space for combined samples |
| $X_{pr,i_p}$ | – | $i_p$-th number of input space for prior knowledge |
| $X_{test}$ | – | Input space of testing samples |
| $X_{test,q}$ | – | $q$-th number of input space for testing samples |
| $X_p^{test}$ | – | $p$-th dimension of input space for testing samples |
| $y_i$ | – | $i$-th number of output space for training samples |
| $y_{c,i_t}$ | – | $i_t$-th number of output space for combined samples |
| $y_{pr,i_p}$ | – | $i_p$-th number of output space for prior knowledge |
| $y_{test,q}$ | – | $q$-th number of output space for testing samples |
| $\hat{y}_i$ | – | Estimated $i$-th number of output space for training samples |
| $\hat{y}_{test}$ | – | Estimated output space for testing samples |
| $Y$ | – | Output space of training samples |
| $Y_c$ | – | Output space of combined samples |
| $Y_{pr}$ | – | Output space of prior knowledge |
| $Y_{test}$ | – | Output space of testing samples |
| $\hat{Y}$ | – | Estimated of output space for training samples |
| $Z$ | – | Objective space |
| $\theta_i^{(.)}$ | – | Angle of arm robot position |
| $\lambda$ | – | Free parameter that control the generalization of the regressed function |
| $\hat{\lambda}$ | – | Estimated free parameter that control the generalization of the regressed function |
| $\gamma$ | – | Estimated free parameter to estimate threhold value |
| $\|X_k\|$ | – | $L_2$norm of the k-th input space of training samples |
| $\Delta error_i$ | – | The error between $\hat{y}_i$ and ,$y_i$ |
| $\Delta ERR$ | – | The vector space in column vector of all corresponding error |

## LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1 Brief Overview of Small Sample Problems

The application of learning from small samples has gained increasing attention in many fields, such as pulp and paper industry [1], scheduling for manufacturing environment [2], assembly process in semiconductor manufacturing [3], engine control problem [4], and biological activities [5]. In real-world applications, usually, it is difficult to obtain sufficient training samples as it is too expensive [6, 7], not informative and redundant [8, 9]. Since the existing techniques focus on large and sufficient training samples problems, researchers have been searching for various methods to address the limitation of learning from small training samples by filling up the information between samples [10, 11].

There are very few small training samples definitions, yet there is no agreed definition on this matter. Vapnik and Chervonenkis [12] consider the samples is small if the ratio size-of-training-set over Vapnik-Chervonik (VC) dimension is smaller than 20. In general, VC dimension is a measure of how many different kinds of function can be modeled from hypothesis class. The VC dimension is considerably difficult to determine [13]. In algorithmic information theory, the samples are considered small based on the Kolmogorov complexity. The Kolmogorov complexity refers to the length of the strings shortest description in some fixed universal description language. Hence, if the string whose Kolmogorov complexity is small relative to the strings size are defined to have small information content [14]. Andonie [10] in his work defined the training samples are small if the number of samples, $n$, and the dimensionality size of samples, $d$, are comparable. Meanwhile, a few authors did not define the definition of small samples but they simply used several numbers of training samples in building the model. Jang and Cho have defined the adequacy of samples must be greater than 40 based on their proposed technique called Observational Learning Algorithm [15]. The

Diffusion Neural Network (DNN) [6] approximates non-linear function with training samples less than nine. The model with training samples of less than 12 to improve the process modeling in pulp industry is carried out by Robert Lanouette et. al [1]. [16] has proposed a method to approximate high-order non-linear function employed as small as 10 training samples to develop a model. Bloch et. al [4] employed at most 6 experimental data to model the in-cylindrical residual gas fractions in Spark Ignition engine with Variable Camshaft Timing for engine control. With very limited literature in defining the size of small dataset and assuming that all the observed samples are sparse, the training samples with less than 30 is considered regardless number of dimensionality as small for regression problem for this study as given by [17].

In general, there are two branches of research areas when solving small samples problems. The first is classification problems, and the second is regression problems. A classification problem is the problem of assigning data to one of a set of pre-determined categories in which it refers the output of the data is in a discrete form. Meanwhile, the regression problem refers to a method of finding relationship between the input and the output of data where the output is in a continuous form. In other words, a classification problem deals with discrete output and a regression problem deal with continuous output. In general, there is a certain amount of works have been done in the past decade to solve small samples problem for classification problem [18]. However, only few works have been presented in literature when solving regression problem and all the techniques available in solving classification problems cannot be directly implemented within a regression framework [7].

## 1.2 Problem Statements

Insufficient training samples may introduce a conflict between the number of samples and the model complexity [9]. Therefore, insufficient training samples cannot provide enough information as there may exist a gap between samples [16]. Hence, the available training samples do not cover the desired input space where the built model may fail to generalize to new unseen samples.

Most of the studies on this problem focus on finding and choosing a different type of hypothesis's classes, $F$, by using several data-driven techniques such as artificial neural network (ANN), Gaussian process (GP), and support vector regression (SVR). The most important issue when solving small training samples problems is

how the technique can generalize to new unseen samples. However, there is no optimal solution for this problem as it is difficult [5]. For instance, in Tsai and Lis proposed method [16], the technique to find an unknown target function, $f$, is by introducing artificial samples to reduce the number of possible of hypothesis's class. However, relying on ANN as a technique to solve small sample problems may cause a difficult problem in determining the model structure, especially the number of hidden nodes. Moreover, in SVR and GP, the selection of models parameter to appropriately determine the hypotheses classes, $F$, is chosen arbitrarily when solving small training samples problem as the focus of the study only to provide an approach to solve small samples problem.

Incorporating a prior knowledge is a plausible method in facilitating the data-driven techniques to improve the quality of the regression. In general, prior knowledge refers to available information about the problem being solved in addition to the training samples [7]. The incorporation of prior knowledge can be used to limit the hypotheses classes, $F$, when only few training samples is available [19] and also when fast computational time is required [8]. In fact, prior knowledge is present in most research fields such as from domain experts [20], and simulation models [4].

In general, the existing approaches in incorporating prior knowledge rely on one optimal solution which is carried out as a single-objective optimization problem [4, 21, 9]. Hence, it requires the pre-selection of coefficients in formulating the learning function of the regression algorithms. Moreover, most of the existing approaches introduce several variables in the formulated learning function in order to finely obtain a good model. Hence, it requires a non-trivial tuning mechanism to achieve the objective. As a result, the existing approaches may simply ignore the possible consequences of the pre-selection of coefficients as limited knowledge is accessible. In general, the selection of coefficients controls the fitness of the learning function either closely approximates to the training samples or prior knowledge, respectively. As the nature of the prior knowledge is considered as less accurate [8] and bias to prior model [7], relying on a single-objective optimization problem which requires a precise selection of the coefficients is difficult.

In general, there are several questions arose based on the given statement from the previous paragraphs.

1. What techniques can offer a better solution when solving regression problem with insufficient samples?

2. How to incorporate prior knowledge for regression problem based on post-selection approach?

## 1.3 Objectives

The objectives of this research in solving regression problem with small training samples are given as follows:

1. To enhance the existing NWKR to solve the regression problem when the available training samples are non-noisy and insufficient.
2. To improve the technique in (1) in order to solve the regression problem with noisy and limited samples.
3. To incorporate prior knowledge based on Pareto optimality concept using the technique in (1) to solve a complex regression problem with noisy and limited samples.

## 1.4 Scope of Work

In designing and developing the proposed WKR and its extension, the scopes of the research are defined as follows:

1. The regression problem with insufficient sample where the input is in continuous value and its limit only to univariate regression problem. The univariate problem refers to the problem with only one output variable, $v = f(U_1, U_2, ..., U_m)$, where *v* is the only output variable, *f* is the target function and *U*'s are the input variables.
2. The problem is also limited to batch data problem. The batch data problem does not concern on the sequence of samples as compared to time-series data when building the model. Hence, the proposed technique treats the available training samples equally as it is independent from the sequence.
3. In incorporating prior knowledge, the proposed technique only focuses on the approach of incorporation. In general, the proposed technique does not consider

the source of the prior knowledge or how to generate a good and reliable prior knowledge.

## 1.5 Thesis Contribution

The main contribution of this thesis is the enhancement of NWKR, which known as WKR. The proposed WKR comes with the generalized models parameter by inheriting one important feature of NWKR when determining the smoothing parameter. The experiment results, which will be shown and discussed in Chapter 3, are the main evidence in terms of generalization and feasibility of the proposed technique. In general, the main aim of introducing the proposed technique is to establish a new technique which can offer a better generalization capability when solving small training samples problems. However, the WKR suffers from certain disadvantage, especially it only limits to non-noisy sample problems.

The second contribution is the extension of WKR when addressing noisy and small training samples problems. In general, the extension employed some existing features that to be associated in the framework of WKR such as a frame based cross-validation technique, new formulation learning functions, and several learning techniques. This extension is not only focused on solving noisy training samples problems but more important as a technique in initializing the population when introducing a new approach of incorporating prior knowledge.

The third contribution is a new approach in incorporating prior knowledge. In general, the existing approach in incorporating prior knowledge is based on one optimal solution. Hence, the need of precise information in defining the coefficients involved is difficult and critically important. This can be considered as pre-selection approach in which ignoring the consequence of the coefficients selection. In this thesis, a new approach is introduced based on Pareto-optimality concept in multi-objective optimization environment. Hence, a multi-objective optimization technique known as non-dominated sorting genetic algorithm II (NSGA-II) is employed. The entire proposed framework can be considered as post-selection approach. The post-selection involves with the selection of solutions instead of the selection of coefficients. This new technique of incorporating prior knowledge is proved to be less subjective, and it offers more generalized model when solving the small samples problems. Also, three challenges when implementing the new approach of incorporating prior knowledge is

highlighted. Hence, the contribution is further scrutinized into three small divisions, which include (1) the technique to initialize the population (2) the technique to address uncertainty and (3) the technique to avoid over-fitting. The technique also could be further research especially in terms of the selection the best solution.

## 1.6 Thesis Organization

This thesis is organized as follows. The literature review of the problem in regression with small samples and the employed techniques is placed in Chapter 2. Chapter 3 focuses on the development and experiment of WKR for solving non-noisy and small samples problems. On the other hand, the extension of WKR with several conducted experiments for solving noisy and small samples problems is explained in Chapter 4. Chapter 5 explains the concept of Pareto optimality in a view of new technique when incorporating prior knowledge in multi-objective environment. Chapter 6 ends this thesis with conclusion as well as some research directions based on this research. Finally, the references are placed at the back of this thesis.

## REFERENCES


1. Lanouette, R., Thibault, J. and Valade, J. L. Process modeling with neural networks using small experimental datasets. *Computers & Chemical Engineering*, 1999. 23(9): 1167–1176.

2. Li, D., Chen, L. and Lin, Y. Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research*, 2003. 41(17): 4011–4024.

3. Lee, W. and Ong, S. Learning from small data sets to improve assembly semiconductor manufacturing processes. *2nd ICCAE 2010*. Singapore. 2010. 50–54.

4. Bloch, G., Lauer, F., Colin, G. and Chamaillard, Y. Support vector regression from simulation data and few experimental samples. *Information Sciences*, 2008. 178(20): 3813–3827.

5. Andonie, R., Fabry-Asztalos, L., Abdul-Wahid, C., Abdul-Wahid, S., Barker, G. and Magill, L. 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. 8(1): 80–93.

6. Huang, C. and Moraga, C. A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, 2004. 35(2): 137–161.

7. Lauer, F. and Bloch, G. Incorporating prior knowledge in support vector regression. *Machine Learning*, 2008. 70(1): 89–118.

8. Leary, S. J., Bhaskar, A. and Keane, A. J. A knowledge-based approach to response surface modelling in multifidelity optimization. *Journal of Global Optimization*, 2003. 26(3): 297–319.

9. Sun, Z., Zhang, Z. and Wang, H. Incorporating prior knowledge into kernel based regression. *Acta Automatica Sinica*, 2008. 34(12): 1515–1521.

10. Andonie, R. Extreme Data Mining: Inference from Small Datasets. *International Journal of Computers, Communications & Control*, 2010. 5(3):

280–291.

11. Zhu, R., Zhang, L. and Chen, A. A new method to assist small data set neural network learning. *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on*. IEEE. 2006, vol. 1. 17–22.

12. Vapnik, V. *The nature of statistical learning theory*. springer. 1999.

13. Matoušek, J., Welzl, E. and Wernisch, L. Discrepancy and approximations for bounded VC-dimension. *Combinatorica*, 1993. 13(4): 455–466.

14. Balcázar, J. L. and Book, R. V. Sets with small generalized Kolmogorov complexity. *Acta Informatica*, 1986. 23(6): 679–688.

15. Jang, M. and Cho, S. Observational learning algorithm for an ensemble of neural networks. *Pattern analysis & applications*, 2002. 5(2): 154–167.

16. Tsai, T. and Li, D. Approximate modeling for high order non-linear functions using small sample sets. *Expert Systems with Applications*, 2008. 34(1): 564–569.

17. Huang, C. Information diffusion techniques and small-sample problem. *International Journal of Information Technology & Decision Making*, 2002. 1(2): 229–249.

18. Lauer, F. and Bloch, G. Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing*, 2008. 71(7): 1578–1594.

19. Niyogi, P., Girosi, F. and Poggio, T. Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, 1998. 86(11).

20. Liao, W. and Ji, Q. Learning Bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, 2009. 42(11): 3046–3056.

21. Mangasarian, O., Shavlik, J. and Wild, E. Knowledge-based kernel approximation. *The Journal of Machine Learning Research*, 2004. 5: 1127–1141.

22. Rumelhart, D. E., Hinton, G. and Williams, R. Learning Representations by Back-Propagating Errors. *Nature*, 1986. 323: 553–536.

23. Hagan, M. T. and Menhaj, M. B. Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on*, 1994. 5(6): 989–993.

24. Latha, P., Ganesan, L. and Annadurai, S. "Face Recognition Using Neural

Networks,". *Signal Processing: An International Journal (SPIJ), CSC Journals, Kuala Lumpur, Malaysia*, 2009. 3(5): 153–160.

25. Lu, S. and Basar, T. Robust nonlinear system identification using neural-network models. *Neural Networks, IEEE Transactions on*, 1998. 9(3): 407–429.

26. Su, T., Jhang, J. and Hou, C. A Hybrid Artificial Neural Networks and Particle Swarm Optimization for Function Approximation. *International Journal of Innovative Computing, Information and Control*, 2008. 4(9): 24492457.

27. Smagt, P. Minimisation methods for training feed-forward networks. *Neural Networks*, 1994. 7(1): 1–11.

28. Wray, J. and Green, G. Neural networks, approximation theory, and finite precision computation. *Neural networks*, 1995. 8(1): 31–37.

29. Friedman, J. An overview of predictive learning and function approximation. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, 1994. 136: 1–61.

30. John, R. S. and Draper, N. R. D-optimality for regression designs: a review. *Technometrics*, 1975. 17(1): 15–23.

31. Cho, S., Jang, M. and Chang, S. Virtual sample generation using a population of networks. *Neural processing letters*, 1997. 5(2): 21–27.

32. Cho, S. and Cha, K. Evolution of neural network training set through addition of virtual samples. *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*. IEEE. 1996. 685–688.

33. Chen, C.-w., Chen, D.-z., Ye, X.-q. and Hu, S.-x. Feedforward networks based on prior knowledge and its application in modeling the true boiling point curve of the crude oil. *Journal of Chemical Engineering of Chinese Universities*, 2001. 15(4): 351–356.

34. Fletcher, L., Katkovnik, V., Steffens, F. and Engelbrecht, A. Optimizing the number of hidden nodes of a feedforward artificial neural network. *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*. IEEE. 1998, vol. 2. 1608–1612.

35. Lazaro, M., Perez-Cruz, F. and Artes-Rodriguez, A. Learning a function and its derivative forcing the support vector expansion. *Signal Processing Letters, IEEE*, 2005. 12(3): 194–197.

36. Yuan, J., Liu, C., Liu, X., Wang, K. and Yu, T. Incorporating prior model into

Gaussian processes regression for WEDM process modeling. *Expert Systems with Applications*, 2009. 36(4): 8084–8092.

37. Lawson, C. L. and Hanson, R. J. *Solving least squares problems*. vol. 161. SIAM. 1974.

38. Wang, F. and Zhang, Q.-J. Knowledge-based neural models for microwave design. *Microwave Theory and Techniques, IEEE Transactions on*, 1997. 45(12): 2333–2343.

39. Fox, J. *Applied regression analysis, linear models, and related methods*. Sage. 1997.

40. Hardle, W. *Applied nonparametric regression*. vol. 5. Cambridge Univ Press. 1990.

41. Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 2005. 27(2): 83–85.

42. Zhang, J., Huang, X. and Zhou, C. An improved kernel regression method based on Taylor expansion. *Applied Mathematics and Computation*, 2007. 193(2): 419–429.

43. Watson, G. Smooth regression analysis. *Sankhy : The Indian Journal of Statistics, Series A*, 1964. 26(4): 359–372.

44. Hansen, B. E. Lecture notes on nonparametrics. *Lecture notes*, 2009.

45. del Río, A. Q. Comparison of bandwidth selectors in nonparametric regression under dependence. *Computational statistics & data analysis*, 1996. 21(5): 563–580.

46. Schaffer, J. D. *Some experiments in machine learning using vector evaluated genetic algorithms*. Technical report. Vanderbilt Univ., Nashville, TN (USA). 1985.

47. Goldberg, D. E. and Holland, J. H. Genetic algorithms and machine learning. *Machine learning*, 1988. 3(2): 95–99.

48. Zitzler, E., Thiele, L., Zitzler, E., Zitzler, E., Thiele, L. and Thiele, L. *An evolutionary algorithm for multiobjective optimization: The strength pareto approach*. Citeseer. 1998.

49. Osyczka, A. and Kundu, S. A new method to solve generalized multicriteria optimization problems using the simple genetic algorithm. *Structural optimization*, 1995. 10(2): 94–99.

50. Deb, K., Agrawal, S., Pratap, A. and Meyarivan, T. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. *Lecture notes in computer science*, 2000. 1917: 849–858.

51. Silverman, B. W. *Density estimation for statistics and data analysis*. vol. 26. CRC press. 1986.

52. Allen, D. M. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 1974. 16(1): 125–127.

53. Hurvich, C. M. and Tsai, C.-L. A corrected Akaike information criterion for vector autoregressive model selection. *Journal of time series analysis*, 1993. 14(3): 271–279.

54. Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1998. 60(2): 271–293.

55. Zhang, X., Brooks, R. and King, M. A Bayesian approach to bandwidth selection for multivariate kernel regression with an application to state-price density estimation. *Journal of Econometrics*, 2009. 153(1): 21–32.

56. Chen, Y., Gupta, M. R. and Recht, B. Learning kernels from indefinite similarities. *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. 2009. 145–152.

57. Chen, K., Ying, Z., Zhang, H. and Zhao, L. Analysis of least absolute deviation. *Biometrika*, 2008. 95(1): 107–122.

58. Moreno, L., Blanco, D., Muñoz, M. and Garrido, S. L1–L2-norm comparison in global localization of mobile robots. *Robotics and Autonomous Systems*, 2011. 59(9): 597–610.

59. Boyd, S. and Vandenberghe, L. *Convex optimization*. New York, NY, USA: Cambridge Univ Press. 2004.

60. Tikhonov, A. and Arsenin, V. Solutions of ill-posed problems. *WH Winston, Washington, DC*, 1977.

61. Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970. 12(1): 55–67.

62. Ng, A. Y. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004. 78.

63. Friedman, J., Hastie, T. and Tibshirani, R. *The elements of statistical learning*. vol. 1. Springer Series in Statistics. 2001.

64. Engelbrecht, A. *Computational Intelligence, An Introduction*. John Wiley & Sons, England. 2002.

65. Moore, A. W., Hill, D. J. and Johnson, M. P. An empirical investigation of brute force to choose features, smoothers and function approximators. *Computational learning theory and natural learning systems*. Citeseer. 1992. 361–379.

66. John, H. Adaptation in natural and artificial systems. *Ann Arbor: University of Michigan Press*, 1975.

67. Goldberg, D. E., Deb, K. and Clark, J. H. Genetic algorithms, noise, and the sizing of populations. *Complex systems*, 1991. 6: 333–362.

68. Wright, A. H. *et al.* Genetic Algorithms for Real Parameter Qptimization. *FOGA*. Citeseer. 1990. 205–218.

69. Eshelman, L. J. chapter Real-coded Genetic Algorithms and Interval-Schemata. *Foundations of genetic algorithms*, 1993. 2: 187–202.

70. Deb, K. and Agrawal, R. B. Simulated binary crossover for continuous search space. *Complex Systems*, 1994. 9: 1–34.

71. Voigt, H.-M., Mühlenbein, H. and Cvetkovic, D. Fuzzy recombination for the breeder genetic algorithm. *Proc. Sixth Int. Conf. on Genetic Algorithms*. Citeseer. 1995. 104–111.

72. Michalewicz, Z. *Genetic algorithms+ data structures= evolution programs*. springer. 1996.

73. Hansen, N. and Ostermeier, A. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*. IEEE. 1996. 312–317.

74. Deb, K. and Goyal, M. A combined genetic adaptive search (GeneAS) for engineering design. *Computer Science and Informatics*, 1996. 26: 30–45.

75. Milanič, S., Strmčnik, S., Šel, D., Hvala, N. and Karba, R. Incorporating prior knowledge into artificial neural networksan industrial case study. *Neurocomputing*, 2004. 62: 131–151.