Learning Systems
DA8008 (HT12) Halmstad University

Regression Model for cooling in a H2O2 process

Yuantao Fan, Mingkun Yang

1 Introduction

This report is focus on Constructing a model for cooling a H2O2 process. This model control is a non-linear task. The goal with the control is to keep the temperature constant and the fluid in liquid format. In this case, A large amount of cooling is required and how much the valve shall opened need to be calculated. The aim is to construct a model with the all input and output variable to control the valve opening.

The process that generated the data is EKA chemicals Hydrogen Peroxide production. The interval between each measurement is 10 minutes and lastly the data has been decimated to 30 min measurements by averaging over 3 samples. The output is the valve opening of a valve connected to one (out of two) heat exchangers.

In this report, covariance between every inputs and outputs is calculated to rank and find the important feature. A linear model is constructed by using the most important feature which is figure out by calculate the covariance, Cross-validation is implemented to estimate the model generalize ability. Also, data transform is applied and PCA (principle components analysis) is used to extract useful component. Multilayer perceptron is implement to find a best model to handle this problem. Features for MLP input is extracted from original data set by using covariance estimation, forward selection and backward elimination. Important data, model and performance is present in the result.

2 Methodology

In this section, Method implemented for model construct is presented, including Data transformation, covariance and correlation estimation, cross-validation, linear model fitting, Principle component analysis, partial least squares fitting and multilayer percetron learning using feature extraction.

2.1 Data Transformation

For data transformation, Logrithm is used to convert original data format to destination data format, Because the data has a wide distribution according to Y axis, we can dectect some outliers of data. Since linear regression is sensitive to outlier, it's crucial to eliminate the outliers. When apply linear regression with logrithm transform on input data, it's more likely to have a small error compare to original data set. After logrithm, some of the data can be regard as outlier and 13 outliers is eliminated in this case.

2.2 Covariance & Correlation Estimation

Covariance between input and output is calculated to rank the importance of every variable. Correlation coefficient is the normalization of covariance matrix, the most importance input feature is the feature has biggest value of correlation coefficient with respect to the output.

2.3 Cross-validation & Linear Regression

Cross-validation is applied to improve the ability of generalization. In this case, 10 fold validation is used for linear fitting and model estimation. The original sample is randomly partitioned into 10 subsamples. Linear regression is used by inputing high ranked important feature, 10 fold validation is used to estimate the generalize error.

2.4 PCA & PLS

Principle component analysis is applied to extract most important information from input data, according to the proportion of the eigenvalue. Then implement PLS using the most important feature to solve the regression problem. In this case the first principle component has the 98% percentage of total. So this component is picked to do the partial least squares regression. Order from 1 to 76 is tested. The one with lowest error is chosen to compare with other method.

2.5 MLP

Multilayer preceptron is implemented to solve this regression problem. For two input data set mentioned in the project description, (Feature Set A={3,6,49,50,51,52,53,65} Set B={4,10,19,27,28,38,42,50}), 3 layer neural network (MLP with one hidden layer) is constructed. 1 to 80 hidden node is tested. Node number with smallest error is chosen to test the model performance.

3 Data

In this regression problem, data consists 3 matrix: XtrainDS (4466x65), YtrainDS (4466x1) and XtestDS (2971x65). All input variable of the process is contained in XtrainDS and output matrix YtrainDS present the output value, which is the valve opening.

First we plot the data according to the order of importance calculated by correlation coefficient. For highest 4 feature: #10,9,65,37, Fig 3.1 shows the result:

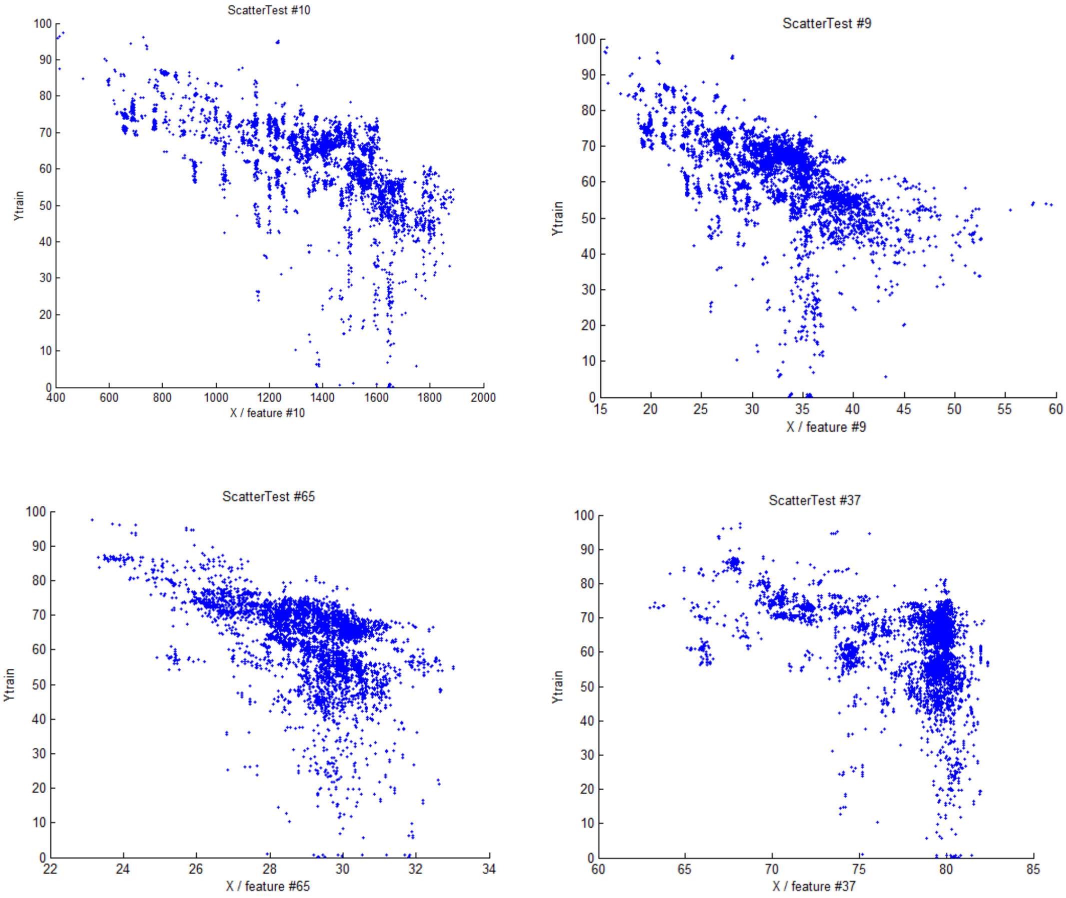

Figure 3.1 Input Feature #10,9,65,37 with respect of Y

For important feature shows in figure 3.1, It seems the data is distributed largely along Y axis, thus it's detrimental for linear regression that we need to model in this case. So logrithm is applied to transform data so that it's easy to find outlier and it's helpful to generate a linear model. Figure 3.2 shows the result of logrithm on feature #10.

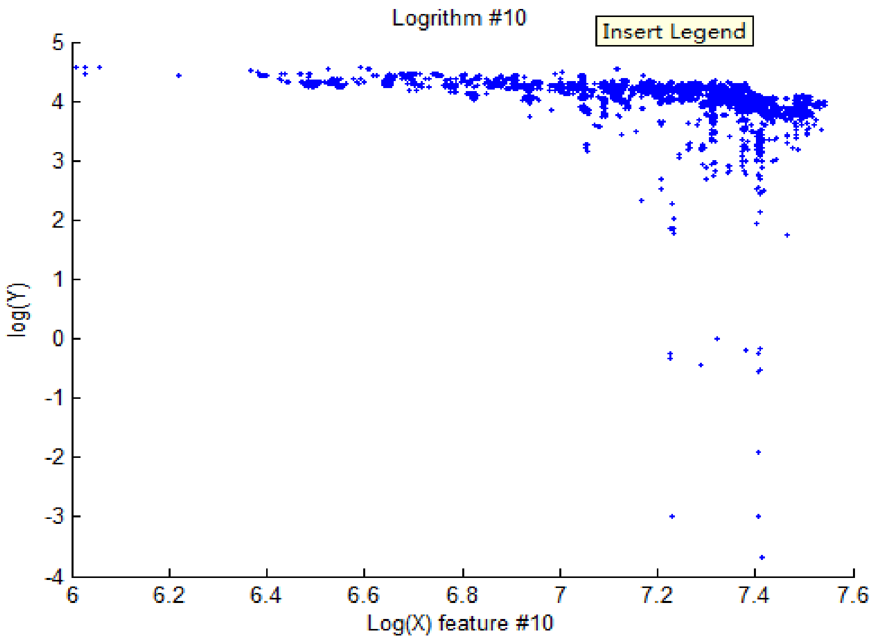

Figure 3.2 Logrithm of Feature #10 and Y

The result show data is more compact according to Y axis. And there is still some outliers, it's also detrimental for linear regression. Then the point that smaller than 1 will be removed. The result shows on figure 3.3(Feature #10).

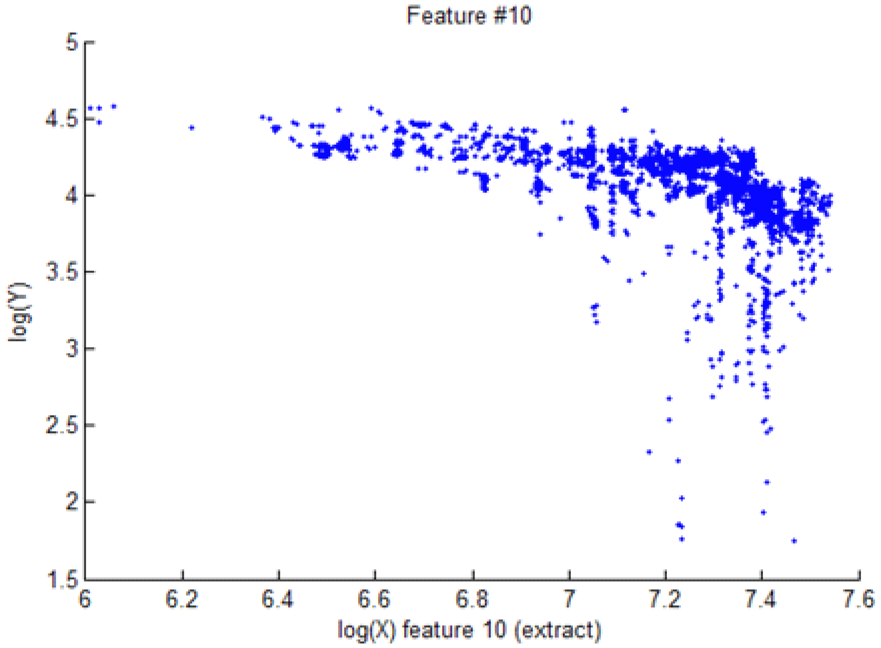

Figure 3.3 Feature #10 (outlier rejected)

Logrithm using here is aim to detect the outliers. The rest of models will use original data format. We calculate the correlation coefficient of every feature, the result shows that feature #10, #9, #65, #37, #8 is ranked as high 5 important feature.

4 Result

In this section, result of models including Linear regression, PCA and PLS, MLP is presented. Cross validation is applied on these technique.

4.1 Linear Regression

For linear regression, Cross validation is used to improve and estimate the generalize ability of the model. The most important ranked high five feature: #10, #9, #65, #37, #8 is used as input. Data transform is used here, we enumerate the combination of 5 feature, and put them into regression algorithm. The result shows that it's better when you apply logrithm on input data.

First input the original data without data transform, the result calculated shows MSE=99.5331 when using five feature #10, #9, #65, #37, #8 for input. If we apply logrithm on inputs, the result shows that MSE = 90.7031 when using feature #9, #65, #37 #10 #8 as input. So for our best linear model Y=B*X, B = [-40.0674 -62.2341 98.2964 0.3195 -3.3369], when use feature #9, #65, #37 #10 #8 as input.

4.2 PCA

For principle component analysis, Function princomp() is used, the function returns the matrix of principle component. Calculate the importance of every column according to eigenvalue, the first component has 98% of all. So the first component is used as input for partial least squares regression.

4.3 PLS

For partial least squares regression, the first component is used as input, function poltfit() is used. Order from 1 to 50 is tested and MSE for every model is calculated, the result is shown in figure 4.2, where the highest order (50) has the lowest MSE.

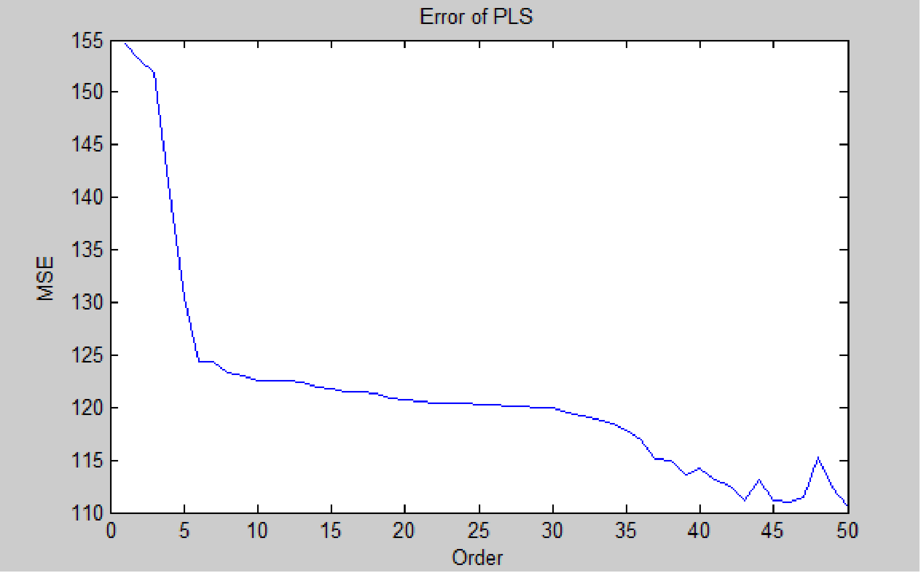

Figure 4.2 Linear Regression using feature #9

For order 50, MSE = 110.565, Fitting result is shown in figure 4.3. Where green circle is the model, blue plus is data.

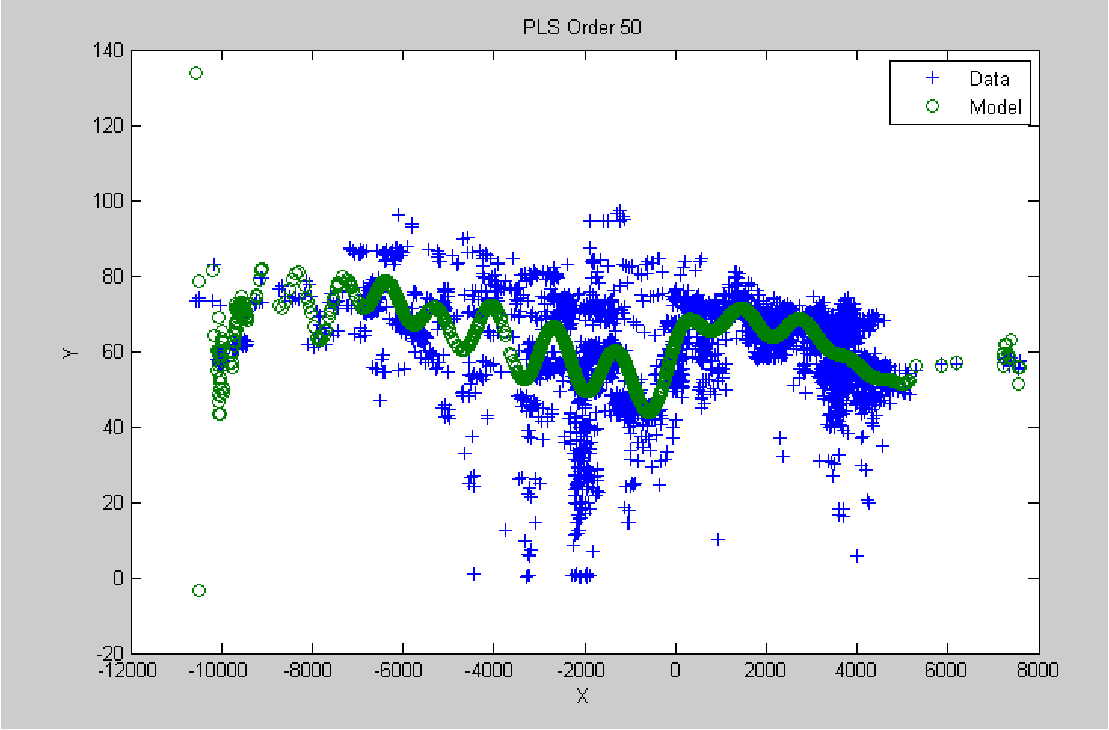

Figure 4.3 PLS Regression

4.4 MLP

For multilayer perception, Different construct is modeled by input different combination of feature. With variables set1 {3, 6, 16, 49, 50, 51, 52, 53, 65}, 1 to 80 hidden node is tested, result is shown in figure 4.4, number of hidden layer with respect to validation error. Node number of 38 has the lowest validation error and

MSE = 43.2054.

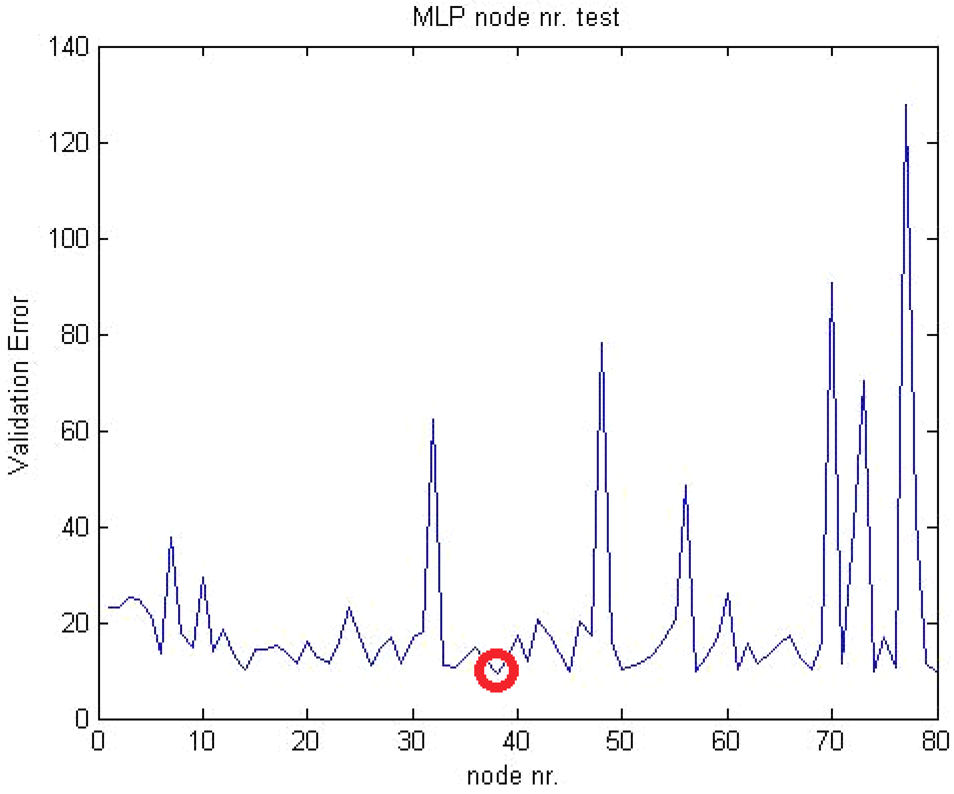

Figure 4.4 MLP node number test for variable set 1

For variable set 2:{4, 10, 19, 27, 28, 38, 42, 50}, 1 to 80 nodes is tested, the result is shown in figure 4.5, 56 node number has the lowest validation error and MSE =48.4743.

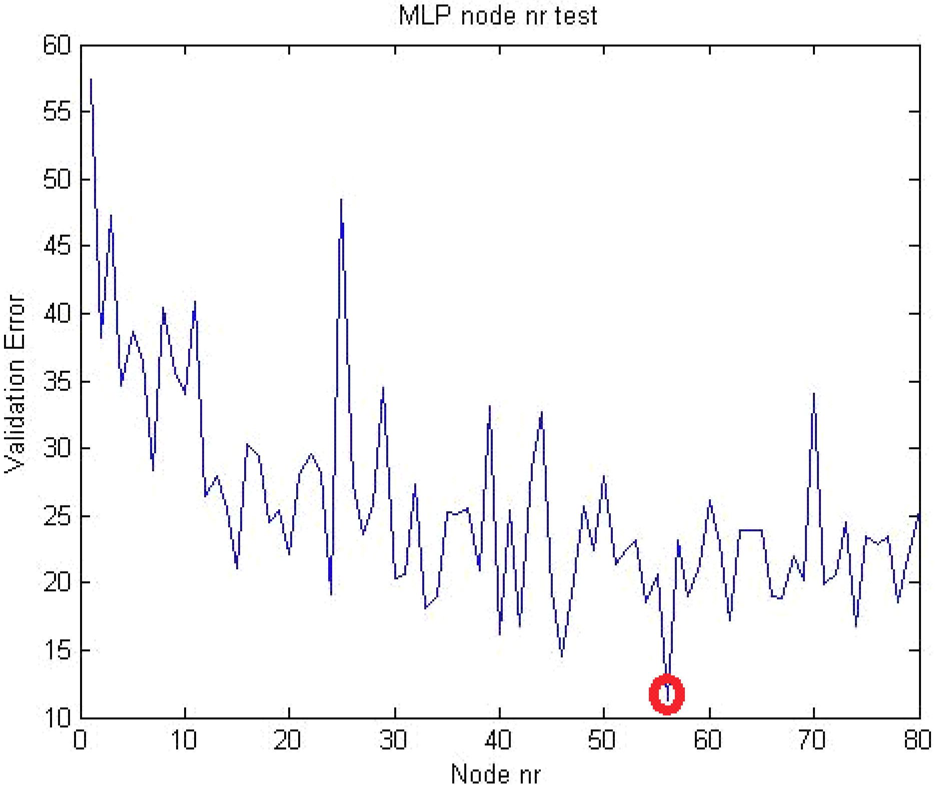

Figure 4.5 MLP node number test for variable set 1

For these two models, Forward selection and backward elimination is used to find a better model. According to the rank of correlation coefficients. For Set 1, we try to eliminate feature #51 and #52 which is ranked very low of correlation coefficients. But it turns out that the mse go bigger. So we readd this 2 features and add 4 more features #9, #10, #37, #8 which is highly ranked in correlation coefficients. 10 hidden nodes has been chosen and MSE is 9.2611. It's quite low, figure 4.6 and 4.7 shows the result of this customized MLP.

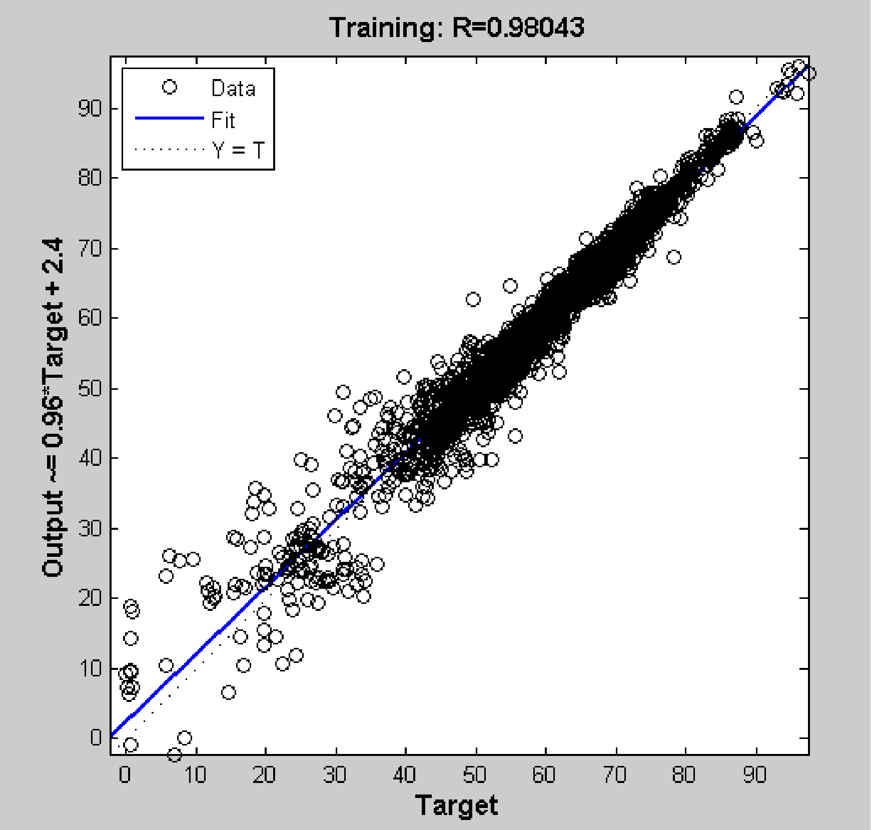

Figure 4.6 Result of customized MLP

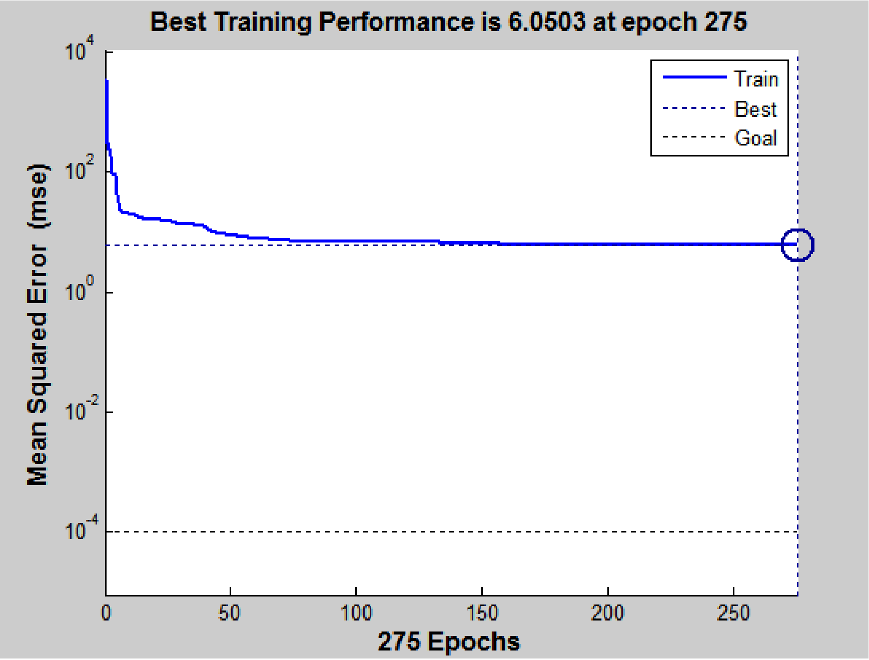

Figure 4.7 Performance of customized MLP