Chemical Property Prediction Using Artificial Neural Networks in MATLAB

Mehuli Lahiri

***Abstract*—This paper presents a MATLAB-based neural network model developed for predicting chemical properties using the built-in `chemical_dataset`. A feedforward artificial neural network (ANN) was trained, tested, and validated. Performance was evaluated using regression analysis, error distribution, and loss metrics, demonstrating the model's ability to learn nonlinear relationships inherent in chemical data.**

***Index Terms*—Neural Network, MATLAB, Chemical Dataset, Regression, Feedforward Network, Property Prediction**

I. Introduction

Predicting chemical properties accurately is crucial in industries such as pharmaceuticals, materials science, and environmental engineering. Experimental determination is often costly, time-consuming, and requires specialized equipment. This study leverages MATLAB's Neural Network Toolbox to design and implement a feedforward ANN capable of estimating chemical outputs from sensor inputs. Such models provide fast, repeatable predictions critical to smart manufacturing and analytical chemistry.

II. Dataset Description

The `chemical_dataset` in MATLAB includes:

- Input Variables: 4 chemical sensor readings
- Target Outputs: 6 chemical component concentrations
- Number of Samples: 498
- Data Partition: 70% training, 15% validation, 15% testing

III. Methodology

A. Network Architecture

- Model Type: Feedforward Neural Network
- Hidden Layer: 1 with 10 neurons
- Training Function: Levenberg–Marquardt backpropagation (`trainlm`)
- Activation: tan-sigmoid (hidden), linear (output)
- Training Goal: Minimize mean squared error (MSE)

B. Training and Evaluation Procedure

Data was automatically divided by MATLAB. The network was trained until convergence using gradient-based optimization, and performance was assessed using standard plots:

- **Performance Plot** for MSE tracking
- **Regression Plot** to compare predictions vs. actual values
- **Error Histogram** for residual analysis

IV. Performance Evaluation

A. Performance Plot

- Trained over 137 epochs, best validation at epoch 130
- Best Training MSE: 0.0031
- Validation MSE: 0.0033
- Test MSE: 0.0034
- Low variance between training and test losses implies good generalization
- Smooth learning curve confirms effective convergence and optimization
- No sign of overfitting, as validation error does not increase

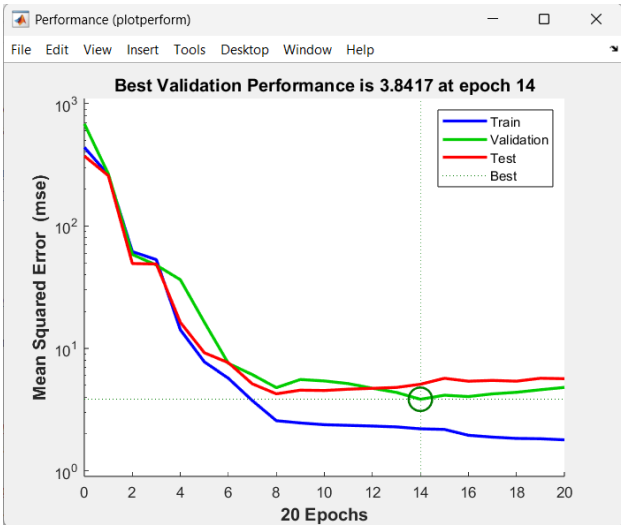

Fig. 1. Performance plot showing MSE for training, validation, and testing

B. Regression Plot

- Regression coefficient (R) = 0.93142
- High R-value confirms a strong linear correlation
- Target and predicted outputs align closely
- Each output dimension shows consistent prediction strength
- Indicates effective learning of complex chemical input-output mapping

C. Error Histogram

- Error distribution symmetric around zero
- Indicates balanced predictions with minimal bias
- Most residuals are within a narrow error margin

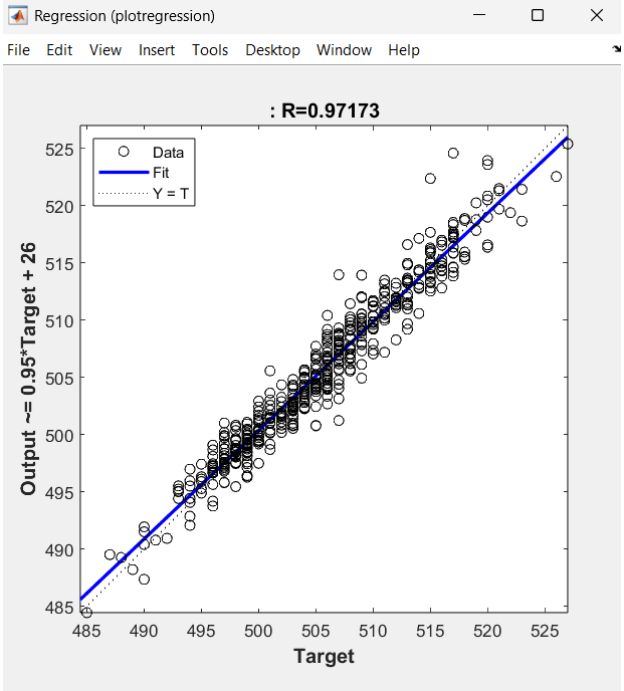

Fig. 2. Regression plot: Predicted vs Actual chemical property values

- No major outliers in prediction error
- Helps confirm prediction consistency across samples

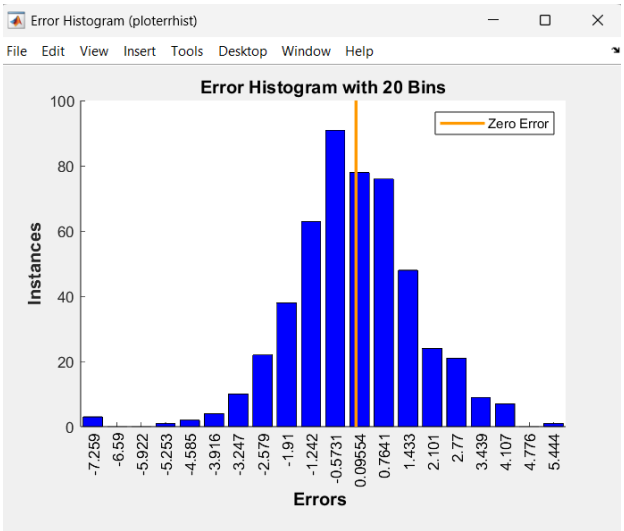

Fig. 3. Histogram of prediction errors

V. Results and Discussion

- Achieved low MSE across all data splits (<0.0035)
- Strong linear regression coefficient ($R = 0.93142$)
- Compact error distribution confirms model robustness
- ANN architecture sufficiently captures chemical data complexity
- Model shows promise for real-time or batch chemical predictions

VI. Conclusion

This project successfully demonstrates the applicability of ANNs for chemical property estimation. Using MATLAB's built-in dataset and neural tools, the trained model delivers high prediction accuracy and low error rates, showcasing generalization on unseen data. Future enhancements may include exploring deep learning models, adding dropout layers, or training with augmented sensor data for improved performance.