A Comparison of Multilayer Perceptron and Support Vector Machines for Wind Speed Prediction

Ciaran Grant *Ciaran.Grant@city.ac.uk*
James Hooper *James.Hooper@city.ac.uk*

April 9, 2018

Abstract

This paper aims to present a detailed comparison of Support Vector Machines (SVM) and Multilayer Perceptron (MLP) used via supervised regression to predict Wind Speed. The orders of each model will be varied from two to ten and using Grid Search Hyperparameter Optimization with 10-fold Cross Validation, the best SVM and MLP is selected for each order. The Mean Squared Error (MSE) and Residuals will be used to compare the models. For this task the SVM with order nine was the best model.

1 Introduction & Hypothesis

Due to the depletion of non-renewable energy sources such as coal, oil and natural gas [1] [3], there is a need to invest in more sustainable sources of energy. Wind powered energy is one such source that is considered a green, low emission and affordable option. China has started to revolutionize their energy systems; they are the country with the largest emission of greenhouse gases due to being the biggest coal-consuming and coal producing nation. It is perhaps unsurprising that a change was needed and they are now the biggest wind power market in the world[5].

The key to efficient use of wind power is the ability to know when and where the wind speeds will be highest and build wind farms in these locations and at these times, a task that is possible using machine learning neural nets [10]. Another important use of short term predicted Wind Speeds is the ability to have engineers on standby if high wind speeds are forecast in order to solve any problems that may arise. Using average daily wind speeds, we will compare the performance of Multilayer Perceptrons (MLP) and Support Vector Machines (SVM) in predicting the speeds of future wind speeds.

Since time series data is being used, the order of the model becomes an extra hyperparameter in the model selection and so, in line with M. Mohandes 2004 paper on 'Support Vector Machines for Wind Speed Prediction' [6], it makes sense to hypothesise that:

1. the optimized SVM will have a lower Mean Squared Error (MSE) than the optimized MLP for each order of input days.
2. the higher order our models, the lower the MSE will be for each model.

2 Data

2.1 Dataset

Meteorology data from the UK Environmental Change Network terrestrial sites will be used in each Neural Network [8]. Data is collected by Automatic Weather Stations (AWS) at each site and verification steps are made such as numeric range checks, categorical checks and logical integrity checks. Hourly summaries from 5-second samplings is collected by the AWS. The available wind speed (m/s) data covers from 1994 to 2015 across 12 different sites around the UK. For the purposes of this report, mean wind speeds for each day were calculated from the Glensaugh site (Site code T02) with the aim of predicting the mean wind speed for the next day. There are 7756 days of wind speeds recorded.

2.2 Initial Analysis of Data

In order to most effectively predict future wind speeds, the data needed to be investigated and processed. To process the data so that it was appropriate for the Neural Networks it was normalised between $[0, 1]$ using:

$$x_{processed} = \frac{x - x_{min}}{x_{max} - x_{min}},$$

where in this case $x_{max} = 16.4208$ and $x_{min} = 0$.

In Figure 1 (Left), it is clear to see that there is a shift in the amplitude of the time series after 2005. This appears to be the same at other sites and is the result in a change of measuring apparatus. In Figure 1 (Right), The distribution helps to point out that the data pre-2005 is skewing the normalised wind speeds closer to zero, so it is worth keeping that in mind when it comes to the results of each model.

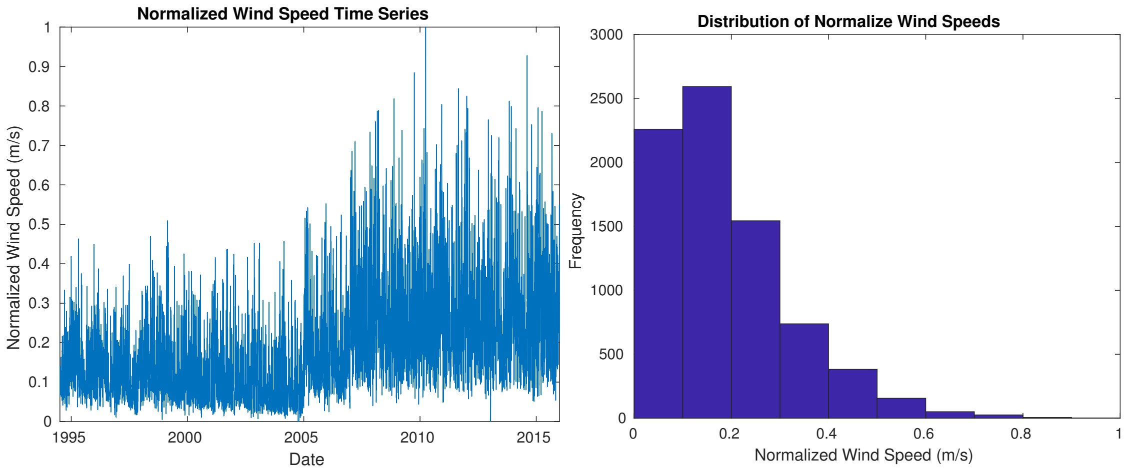

Figure 1: (Left) Normalized Wind Speed Time Series from 1994 to 2015, (Right) Histogram of Normalized Wind Speeds showing the distribution.

3 Neural Networks

3.1 Multilayer Perceptron (MLP)

Multilayer perceptrons consist of an input layer, an output layer and any number of hidden layers in between. The aim of the Multilayer Perceptron is to approximate an underlying function of the input data such that

$$g(x) = \sum_j \alpha_j h(W_j x + \theta_j)$$

approximates the underlying function f(x). Where α_j is the learning rate, W_j are the weights, h(x) is the activation function and θ_j is the bias. Learning involves changing the weights of the neurons via Backpropagation. Backpropagation compares an obtained output with the desired output and sends the error backwards through the network, you can then change the initial weights accordingly and repeat the process, stopping when certain criteria are met or there is no change to the network.

- Advantage: Mimics and exceeds learning that humans are capable of.
- Advantage: Is an efficient universal approximator.
- Disadvantage: Has trouble dealing with multiple local minima of the error function.

3.2 Support Vector Machines (SVM)

The basic idea of Support Vector Machine Regression is to find a hyperplane that best represents the target data. To do this some tricks and hyperparameters are used. Firstly the Kernel trick takes the data to a new dimension so that it is easier to find the hyperplane with the best fit. The hyperparameter, epsilon, is chosen as a tolerance around the hyperplane so that any value that falls within the boundary is not penalised. A box constraint is introduced to 'determine the trade-off between the flatness of the hyperplane and the amount up to which deviations larger than epsilon are tolerated' [9]. Some advantages and disadvantages of the SVM model are as follows:

- Advantage: The kernel trick speeds up computation time considerably.
- Advantage: The nature of the method means a global minimum is always found, the same cannot be said about the MLP.
- Disadvantage: Can be prone to overfitting [2].

4 Training & Evaluation Methodology

As mentioned before, time series data provides an extra hyperparameter for each model as the order of the model also needs to be optimized. Since each order will have different optimal hyperparameters, the process of hyperparameter optimization will be repeated 9 times, where 9 is the number of different orders being tried from $2 - 10$. For every different order, Wind Speed for only one day was predicted due to simplicity. To create data with different input orders, the data was split into rows of $2 - 10$ input days with 1 output day. These rows will all be consecutive days of data where the output day is the day immediately following the input days. No day from the initial time series was used more than once in the data. The data was split into training and test data using the holdout method, where 80% was used for 10-fold Cross Validation and 20% was kept separate to be used as a final test set for the selected model.

For model selection a grid search was used (discussed in more depth in the following section). For all permutations of hyperparameters the MSE was calculated utilising 10-fold cross validation for both SVM and MLP. For each order of input days the minimum MSE was found and the corresponding hyperparamters were stored. These hyperparamters were used to train a final model for each order using all of the training data. Then the MSE (which is the performance measure used throughout this paper) was calculated using all the test data that was initially held back and all the training data.

N.B. there will be a model trained for each order with a set of hyperparameters and an MSE directly relating to that order. The best model will be chosen from these nine using the MSE from the training data as a comparison.

5 Parameter Tuning

To optimize the hyperparameters (Order, Kernel, Epsilon and Box Constraint) for the SVM, a grid search was used to calculate the validation MSE for each permutation of hyperparameters. For each set of hyperparameters 10-fold Cross Validation was used to calculate the validation MSE.

For the Order: $2-10$ were used and the lowest validation MSE was of Order 9, which is nearly in line with our hypothesis that a higher order would improve accuracy.

For the Kernel: Linear, Gaussian and Polynomial were searched through. Each order of the polynomial kernel, performed worse than the Linear and the Gaussian, which produced much more similar results as can be seen in Figure 2 (Right), with the Linear Kernel having a marginally better performance.

For Epsilon: values between 0.01 to 0.5 were searched through, and as can be seen again in Figure 2 (Right), the smallest values of epsilon performed the best.

For the Box Constraint: values between e^{-3} to e^{3} were searched through with little benefit as there was almost no change between values.

Order	Kernel	Epsilon	Box Constraint
9	Linear	0.05	$e^0 = 1$

Table 1: Best SVM Hyperparameters.

To optimize the hyperparameters (Order, Hidden Layers and Initial Learning Rate) for the MLP, a grid search was used to calculate the validation MSE for each permutation of hyperparameters. For each set of hyperparameters 10-fold Cross Validation was used to calculate the validation MSE. Levenberg-Marquardt backpropogation was used in training due to vastly increased computation times.

For the Order: $2-10$ were used and the lowest validation MSE was of Order 9, which is nearly in line with our hypothesis that a higher order would improve accuracy.

For Hidden Layers and Initial Learning Rate: Figure 2 (Left) represents the Grid Search for Order 9 between Hidden Layers and Initial Lerning Rate. Values for Initial Learning Rate were between 0.001 and 1. Values for Hidden Layers were between 5 and 100.

Order	Hidden Layers	Epsilon
9	65	0.001

Table 2: Best MLP Hyperparamters.

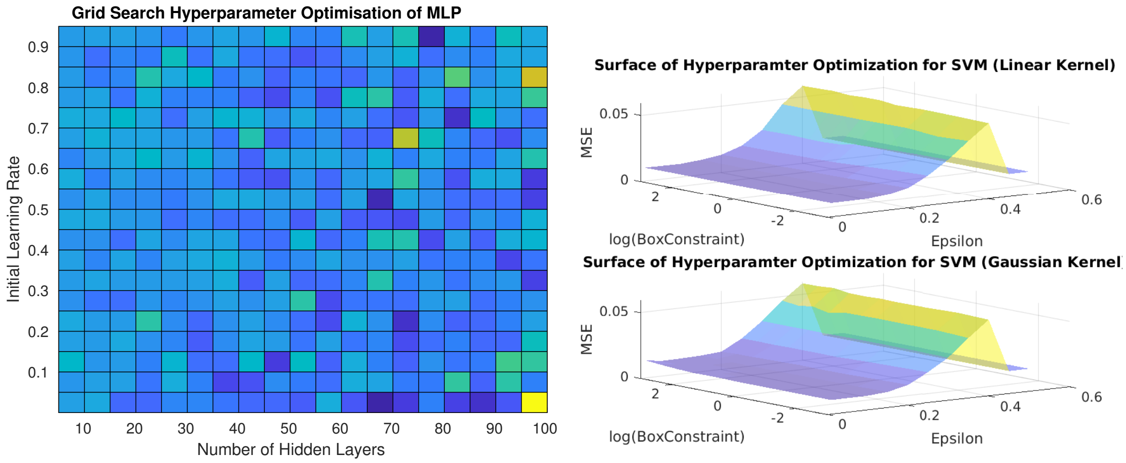

Figure 2: (Left) MLP Grid Search, (Right) SVM Grid Search for both Linear and Gaussian Kernel.

6 Analysis & Critical Evaluation

In Figure 3 (Left), Test MSE is plotted for each model order. In model selection an order of nine was chosen for each model, which is interesting as when those models were used to gain an MSE on the Test Set it is not the smallest MSE. In fact the MSE is increasing as the Order increases, which is in contrast to our second hypothesis. The first hypothesis appears to hold true though as the SVM has a better Test MSE for every Order apart from seven. Taking Ockham's Razor into account it may actually be more appropriate to take the model of order two as it is the simplest model.

When optimizing the hyperparameters for the MLP an order of nine was chosen. Interestingly this order provides the worst Test MSE. Similarly for the SVM an order of nine was chosen and this provides on of the worst test MSE's. This is likely to do with the split of the Training and Test set.

In Figure 3 (Right), the Distribution of Residuals for SVM and MLP are shown on histograms. This shows that the SVM predicts values that are consistently closer to the actual value than the MLP. The MLP is skewed to predicting values that are larger than the actual value, with some outliers around 0.4 away from the actual value.

Knowing what the model will be used for is vitally important as some of the issues highlighted in the plot of residuals may be negated. If the model is being used to choose a site to build Wind Farms, then it is much worse for errors to overestimate the Wind Speed like in the MLP and it is more prudent to choose the SVM as the range of errors is smaller and less likely to overestimate wind speeds.

However if the model was being used by engineers that needed to know if there is likely to be wind speeds so high that the wind farms cannot tolerate them, it would be preferable to overestimate and so choosing the MLP for this task could be more useful.

7 Conclusion

It is now possible to answer the initial hypotheses. For hypothesis one, going on the Test MSE alone the SVM is a better model for predicting wind speeds than the MLP for all orders except seven. The Test MSE increases as order increases, which is contrary to hypothesis. Some future work to consider is:

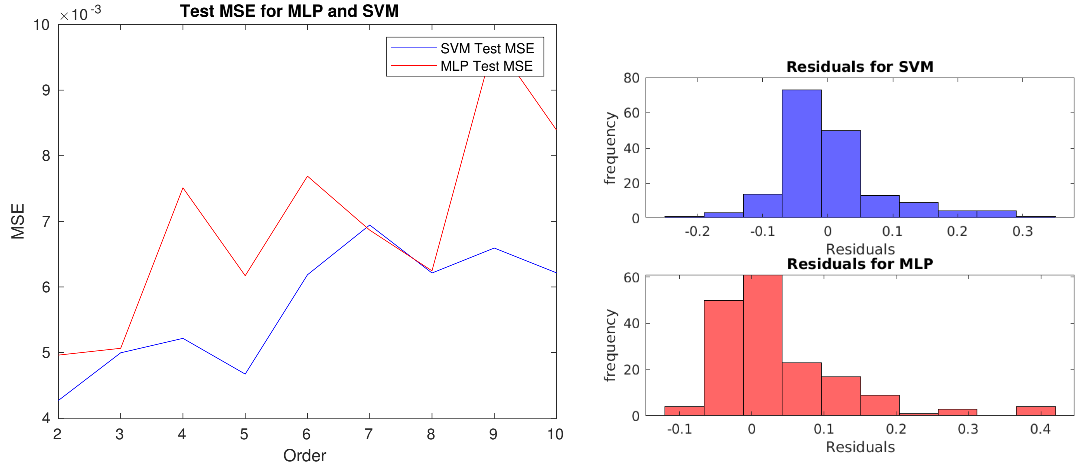

Figure 3: (Left) How the order of the model changes the MSE , (Right) Residuals for the SVM and MLP.

- potentially removing any data before 2005 as it seems to be significantly different to data after 2005.
- wind speed data for hours in the day was available so it would be interesting to use more granular data and see if the accuracy changed at all.
- using a different hyperparameter optimization method like random search or bayesian optimization.
- using the RMSE as the error measure as the number of rows in the data changes with each order and it would likely cause the RMSE to reduce as the Order increases.

References

[1] Capellán-Pérez, I., Mediavilla, M., de Castro, C., Carpintero, Ó. and Miguel, L. (2014). Fossil fuel depletion and socio-economic scenarios: An integrated approach. Energy, 77, pp.641-666.

[2] Cawley, G. and Talbot, N. (2010) Over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research, 2010. Research, vol. 11, pp. 2079-2107.

[3] Höök, M. and Tang, X. (2013). Depletion of fossil fuels and anthropogenic climate change—A review. Energy Policy, 52, pp.797-809.

[4] Jung, J. and Broadwater, R. (2014). Current status and future advances for wind speed and power forecasting. Renewable and Sustainable Energy Reviews, 31, pp.762-777.

[5] Lewis, J. (2012). Green Innovation in China: China's Wind Power Industry and the Global Transition to a Low-Carbon Economy by Joanna I. Lewis. China Review International, 19(4), pp.623-625.

[6] Mohandes, M., Halawani, T., Rehman, S. and Hussain, A. (2004). Support vector machines for wind speed prediction. Renewable Energy, 29(6), pp.939-947.

[7] Mustafidah, H., Hartati, S., Wardoyo, R. and Harjoko, A. (2014). Selection of Most Appropriate Backpropagation Training Algorithm in Data Pattern Recognition. International Journal of Computer Trends and Technology, 14(2), pp.92-95.

[8] Rennie, S. (2017). UK Environmental Change Network (ECN) meteorology data: 1991-2015. NERC Environmental Information Data Centre. https://doi.org/10.5285/fc9bcd1c-e3fc-4c5a-b569-2fe62d40f2f5

[9] Smola, A. and Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), pp.199-222.

[10] Tascikaraoglu, A. and Uzunoglu, M. (2014). A review of combined approaches for prediction of short-term wind speed and power. Renewable and Sustainable Energy Reviews, 34, pp.243-254.