Prediction on Concrete Compressive Strength and Setting Time

Zelin Han

December 18th, 2021

Abstract

Concrete is one of the most important building materials in modern construction and how to get qualified concrete is the key to ensuring building safety. Moreover, to reduce costs, it's also worth studying that how to create concrete using the least time while ensuring quality. Therefore, this project is going to give a prediction of cement which three goals.

The first goal is to predict the concrete compressive strength using 7 ingredients of concrete and setting time. The second goal is to predict the setting time according to the required strength and compositions of the concrete. Finally, this project will give the most important features of making concrete.

The dataset used in the study comes from Pro. I-Cheng Yeh and it's included in UCI Machine Learning Repository named Concrete Compressive Strength Data Set. And the project used three machine learning algorithms to train the prediction models which are PLS (partial least squares regression), KNN (K-nearest neighbors regression), and SVR (support vector regression). Besides, it also used PCA (principal component analysis) and SFS (sequential feature selection) to preprocess and analyze the datasets.

Introduction

As technology advanced and aesthetic standards improved, buildings began to become more and more distinctive. Both the shape and the height of the building are making building standards higher and more difficult. Such high requirements for architecture require not only the design of engineers but also the high performance of building materials. And Concrete is one of the most critical building materials to ensure safety.

Concrete is a composite material made by gel material, aggregate, and water in an appropriate proportion and then hardened over a certain period of time. It is the world's largest use of artificial civil engineering and construction materials due to its hardness, high compressive strength, durability, wide range of sources, simple production methods, low cost, plasticity, and suitability for all kinds of natural environments. [1]

The concrete ingredients include cement, blast furnace slag, water, superplasticizer, and aggregate. Over time, the concrete solidifies to strengthen the building. And the strength of concrete is a highly nonlinear function of setting time and compositions. [1] So, it's hard to get the compressive strength of concrete through simple calculations.

Therefore, the project used machine learning methods to train models that predict the compressive strength by 7 ingredients of concrete and the setting time. During the construction of buildings, workers can judge the quality of the concrete based on the prediction of the current strength and then plan the next stage of the construction.

In addition, in more application scenarios, builders need to plan the construction process by anticipating the setting time of concrete based on the required compressive strength. Thus, this project also try to provide models that can give predictions of setting time based on the ingredients and required strength.

Finally, in order to eliminate some manual errors, this project will give some of the most important features that affect the compressive strength of concrete. During the construction, builders should pay more attention to those aspects to ensure the quality of concrete.

Database Description

In order to train models, the project used the dataset donated by Prof. I-Cheng Yeh in 2007 and it's included in UCI Machine Learning Repository named Concrete Compressive Strength Data Set. [2] This dataset contains 1030 instances with 9 attributes which are 'Cement', 'Blast Furnace Slag', 'Fly Ash', 'Water', 'Superplasticizer', 'Coarse Aggregate', 'Fine Aggregate', 'Age', and 'Concrete Compressive Strength'.

In this dataset, all the data are numerical; only the 'Age' is integer, and the rest attributes are decimals. Moreover, the dataset has complete data which means there is no missing attribute value.

For each attribute, the measurements are kg per cubic mixture for 'Cement', 'Blast Furnace Slag', 'Fly Ash', 'Water', 'Superplasticizer', 'Coarse Aggregate', and 'Fine Aggregate', day from 1 to 365 for 'Age', and MPa for 'Concrete Compressive Strength'.

Methodology

Firstly, I applied three machine learning algorithms which are PLS (partial least squares regression), KNN (K-nearest neighbors regression), and SVR (support vector regression) to train models by using the original dataset.

For the PLS algorithm, it is a statistical method of finding a linear regression model by projecting the predicted variable (Y) and observed variable (X) into a new space and performing least squares regression on these variables. [3] It is used to find the basic relationship between two matrices, which is an implicit variable method to model the covariance in these two spaces. And it’s particularly suitable when the prediction matrix has more variables than the observation matrix, and when the value of X is multicollinearity. [3] The PLS model has a parameter called components which means the number of features used for the model fitting, and it equals to N-1, when the system has N components and there is no interaction between them.

For the KNN algorithm, it is a nonparametric method that approximates the correlation between independent variables and targets by averaging the observed values in an area of the neighborhood. [4] And the size of the neighborhood needs to be set by the analyst according to some studying and testing.

For the SVR, it gives a decision boundary based on the observation points and fits a hyperplane to pass through as many points in the boundary as possible. [5] For the degree of fitness, it has a parameter C which is inversely proportional to the strength of regularization. A large C gives the model low bias and high variance because it penalizes the cost of false regression a lot, and a small C gives the model higher bias and lower variance. [5] In other words, the C suggests how much the model needs to avoid false regression. It looks like the larger the C, the better the prediction. But the cost of a huge C is that the minimum margin will be very small which sometimes leads to a bad regression depending on the characteristic of the dataset. [5] The figure (Fig.1) shows two examples of classification with low and large C, and it's similar to the regression problem.

To get good prediction models, I used the cross-validation method to find the optimal parameters for those regression algorithms. Cross-validation is a practical statistics method that divides a sample of data into small subsets for analysis and validation. The purpose of cross-validation is to test the performance of the model with new data that has not been used to train the model. It can reduce the problems such as overfitting and selection bias, and evaluate the generalization of the model on different independent datasets. Therefore, by using this method on these three regression algorithms, I can find the optimal parameters for the models by calculating its mean square error between the prediction and observation.

After getting the prediction for both concrete compressive strength and setting time on the original data, I'm going to use PCA (principal component analysis) to preprocess data and then train models for the prediction. PCA is a statistical procedure that forms the basis of multivariate data analysis based on the projection method. The most important use of PCA is to represent multivariable data tables as smaller sets of variables (aggregate indexes) to observe trends, jumps, clusters, and outliers. [6] This overview reveals the relationships between observations and variables and between variables. My purpose is to visualize the trends of compressive strength by using the score plot.

Finally, I used SFS (sequential feature selection) to get the most important features that affect the compressive strength of concrete. The forward selection of SFS adds features to form a feature subset in a greedy fashion. By using the cross-validation on the feature subset, I can get the mean square error by regression algorithm to find the features that mostly affect the compressive strength of the concrete.

Results & Discussion

I used python to preprocess the dataset and give a prediction for the targets. I used PLS, KNN, SVR, PCA, and SFS from the sklearn library.

In the first part, I used the original dataset to train the models of PLS, KNN, and SVR to give predictions. Using PLS to predict compressive strength, I firstly used cross-validation to find the optimal number for the parameter ‘n_components’ from 1 to 9. I used MSE as the score to judge the prediction for different number of components and split the dataset into 10 parts to do the cross-validation 3 times. From the plot (Fig.2), it shows that the MSE decreases when the components increase from 1 to 7; and after 7, the MSE becomes stable. This result matches with the principle of PLS regression that the best number of components equals to N-1 since we have 8 features in the training data. By using 7 components, I used 'PLSRegression’ to train the prediction model by both raw dataset and standardized dataset. The scores are 0.637 for raw data and 0.605 for standardized data. This result shows that the PLS regression is not sensitive to the standardized relationship among the data. Thus, I used raw data to train the model and give a prediction of compressive strength. The mean square error between the prediction and observation is 95.679. I tried to plot the real value vs. predict value of strength in the same diagram to visualize the difference between them. But due to the fluctuation over instances (Fig.3), it's hard to see any relationship between the two lines. So I sorted the data by the compressive strength of the original data and calculated the MSE again to check the correctness of my sorting. After this process, I got the plot (Fig.4) of real value vs. predict value and I can clearly see that the prediction is not good, since the prediction has many outliers among the instances. And I used the same way to predict the setting time and plot the diagram of ‘Age’. From the plot (Fig.5), I can see that the model has a bad prediction when the ‘Age’ should be large. Besides, the MSE is 2701.033 which is too large to say this is a suitable model for setting time prediction. So, I tried to find other regression models to give the prediction.

For using KNN, I also used the similar cross-validation method to find the parameter 'n_neighbors’ of KNN regression. And from the plot (Fig. 6), it clearly shows that when the model uses 3 neighbors, it gives the best prediction with minimum MSE. Therefore, I used 3 neighbors to train the KNN model for both raw data and standardized data. The scores are 0.711 for raw data and 0.731 for standardized data. This means the KNN regression model is more sensitive to the relationship of standardized data. So, I calculated the MSE of KNN prediction on standardized data and the result is 70.657 which is better than the PLS regression.

I plotted the real value vs. predict value; and from the diagram (Fig.7), I can see that even the trends of prediction are closer to the real value line, but there are still a lot of outliers among the cases. As for the prediction of setting time, using cross-validation, the plot (Fig.8) shows that when using 8 neighbors, the KNN regression gives the best prediction on 'Age'. By using this parameter, I got the plot (Fig.9) of prediction on 'Age', and it doesn't show any trends close to the real value. Besides, the MSE is 2597.752 which is also too large to say the KNN is a suitable model for setting time prediction. Thus, I tried to find another regression model for the prediction on both compressive strength and setting time.

For SVR mode, I used cross-validation to find the optimal number for parameter C from 1 to 99, since the larger the C the better the model fit the data. From the plot (Fig.10), I can see that the MSE decreases sharply when the parameter C increase. The reason for not trying larger C is because the cost of a huge C is the minimum margin will be very small; besides, from the plot, the MSE seems stable when C is around 99. Thus, I used 99 as parameter C for the SVR model. And the score values of SVR are 0.752 for raw data and 0.830 for standardized data. From this result, I can conclude that the SVR model is the most sensitive model for the standardized relationship among the data. So I used standardized data to train the SVR model and get the MSE of 44.856 which is much better than the PLS and KNN model. From the real value vs. predict value plot (Fig.11), I can see the difference between the two lines is much smaller than the previous plot, and there are no obvious outliers in this model prediction. Therefore, I considered the SVR with 99 as parameter C is a good model for the prediction of concrete compressive strength. As for the prediction of setting time, the cross-validation plot (Fig.12) also shows a rapid decrease through the increase of C, but the MSE is still too large. From the real value vs. predict value plot (Fig.13), there are still some outliers when ‘Age’ is smaller than 180 and the model cannot predict the trends when ‘Age’ is larger than 180. Besides, even with 99 as parameter C, the MSE is still 2214.080. Thus, the SVR model is also not suitable for setting time predictions.

For the second part, I tried to use PCA to preprocess the dataset in order to visualize the trends of prediction. And since the models cannot predict the setting time in an acceptable plot or MSE, I only studied the prediction of compressive strength using the SVR model in the following work. Firstly, I used the 'PCA' model in the sklearn library to fit the model and then get the ‘explained_variance_ratio_' which is the R-square that describes how much that original data can be explained by each component of PCA. The R-square values for each component are 0.285, 0.177, 0.168, 0.127, 0.119, 0.099, 0.022, and 0.004. From the score plot (Fig.14) of t1 vs. t2, I cannot see any pattern or relationship among the points. This is because the first two components have a small R-square value which means only a little relationship is explained by these two components. So, I tried to use multiple components to train the model and give a prediction on compressive strength. I used 5 components to train the SVR model because the first 5 components have 0.875 as R-square and I thought it’s enough to give it a try. I will also use SFS to form a subset for model training and compare the prediction between the PCA-selected data and SFS-selected data in the next part. Back to PCA, I used 5 components to train the SVR model and the MSE of the prediction is 101.219 which is not good compared with the previous prediction. From the real value vs. predict value plot (Fig. 15), I can see there are many outliers among the instances and the prediction line is much worse than the previous plot

(Fig.11). Thus, I considered the PCA is not suitable for this dataset using SVR to predict concrete compressive strength.

Finally, I used SFS to form a subset of data in order to find the most important features that affect the compressive strength. At first, I wrote a function that uses forward search of SFS to give a sequence of features based on the MSE of the prediction, which means the first feature in the sequence is the most important feature that affects the prediction and the last is the least important. The sequence is 'Age', 'Cement', 'Blast Furnace Slag', 'Water', 'Fly Ash', 'Superplasticizer', 'Fine Aggregate', and 'Coarse Aggregate'. In the beginning, I tried to use the first three features in the sequence to train the SVR model, but the MSE of the prediction is 67.619 which is not close to 44.856 the MSE that used the original data. So, I tried to use the first four features in the model training. And this time I got 46.795 as the MSE which is very close to the 44.856. Therefore, I can conclude that 'Age', 'Cement', 'Blast Furnace Slag', and 'Water' are the most important features that affect the compressive strength. Besides, compared to this result that SFS only used 4 features to obtain a better prediction, the PCA dataset used 5 components but got a worse prediction. This also shows that PCA is not suitable for this problem.

Conclusion & Future Work

To summary the above work, the PLS model is not sensitive to the standardized relationship for training models, but the KNN and SVR modes are sensitive to the standardized data. Besides, the SVR gave the best prediction of compressive strength using 99 as parameter C, and the MSE between the prediction and observation is 44.856 which can be considered as a good prediction.

As for the setting time, none of the models can give a good prediction. This is probably because the relationship between ‘Age’ and other features is highly nonlinear function and extremely complex to predict. Besides, the sequence of features that I got in the last part shows that ‘Age’ is the most important feature which affects the prediction of compressive strength. This indirectly proves that 'Age' is a sufficient and unnecessary condition for ‘Compressive Strength’ in prediction. So I considered that maybe there is another feature needed for the prediction of setting time. After reviewing some standards and papers, I found that based on "ASTM Standard C403", the initial setting time of concrete is measured based on penetration resistance of mortar sieved from concrete and it is considered as the time taken to achieve a penetration resistance of 3.5MPa which is an arbitrary value. [7] In addition, the latest research provides a method that uses ultrasonic pulse velocity to measure the shear wave and then evaluate the setting time. Researchers used longitudinal (P-wave) ultrasonic wave propagation, penetrometer-based setting time, semi-adiabatic calorimetry, and formwork pressure to monitor the setting time of concrete; and the study shows that P-wave tests and calorimetry can be used to monitor the hardening and estimate the setting time of concrete. [8] Therefore, I cannot use Prof. I-Cheng Yeh’s dataset to predict the setting time well, since it needs other features of concrete.

Moreover, I got a ranking on the importance of influencing concrete compressive strength which is 'Age', 'Cement', 'Blast Furnace Slag', 'Water', 'Fly Ash', 'Superplasticizer', 'Fine Aggregate', and 'Coarse Aggregate'. Besides, based on the testing, the first 4 features are the most important that affect the compressive strength. Workers need to pay more attention to these aspects.

In the future, I can find other suitable datasets to predict the setting time. Using some datasets that contain the measurement of penetration resistance of mortar sieved from concrete or the measurement of the shear wave of concrete may help the model to predict an acceptable setting time. This will be of great help to the construction industry and workers.

Citation

[1] Kumar Mehta, & Monteiro, Paulo J. M. (2014). Concrete: microstructure, properties, and materials / P. Kumar Mehta, Ph.D.; Paulo J. M. Monteiro (4th ed.). McGraw-Hill Professional.

[2] I-Cheng Yeh (1998). Modeling of strength of high performance concrete using artificial neural networks. Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808

[3] Febrero‐Bande, Galeano, P., & González‐Manteiga, W. (2017). Functional Principal Component Regression and Functional Partial Least‐squares Regression: An Overview and a Comparative Study. International Statistical Review, 85(1), 61–83. https://doi.org/10.1111/insr.12116

[4] Rogel-Salazar, & Taylor & Francis. (2017). Data science and analytics with Python / Jesus Rogel-Salazar. CRC Press, Taylor & Francis Group.

[5] Parbat, & Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. Chaos, Solitons and Fractals, 138, 109942–109942. https://doi.org/10.1016/j.chaos.2020.109942

[6] Abdi, & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews. Computational Statistics, 2(4), 433–459. https://doi.org/10.1002/wics.101

[7] Piyasena, R. R. C., Premerathne, P. A. T. S., Perera, B. T. D., & Nanayakkara, S. M. A. (2013). Evaluation of initial setting time of fresh concrete. In Proceedings of the 19th ERU Symposium.

[8] Wang, X., Taylor, P., Wang, K., & Lim, M. (2016). Monitoring of setting time of self-consolidating concrete using ultrasonic wave propagation method and other tools. Magazine of Concrete Research, 68(3), 151-162.