Prediction of Concrete Compressive Strength Using Multivariate Feature Extraction with Neurofuzzy Systems

Deok Hee Nam
Engineering and Computing Science, Wilberforce University, Wilberforce, OHIO, USA

Abstract - The proposed work shows how to evaluate the target values applying the efficient data handling methods using various multivariate analyses with reduced dimensionalities. In order to explore the proposed methods, neurofuzzy systems developed by the original data and the reduced data are adapted to estimate the high performance concrete through the concrete compressive strength data. In addition, two different paradigms of the reduced dimensionalities are compared to show the better performance between three extracted features and four extracted features. Finally, various statistical categories are applied to determine the best performance among the applied techniques to evaluate the results through the neurofuzzy systems with the original and reduced data of concrete compressive strength.

Keywords – concrete compressive strength, data mining, dimension reduction, feature extraction, multivariate analysis, neurofuzzy system

1 Introduction

In recent years, researches on the ability of concrete have received increasing attention. Concrete is widely used for construction material since it is cost-effective and able to carry relatively high compressive loads. Nevertheless, for the structure of concrete, it is often vulnerable to micro-cracks, which can cause the durability problems of civil infrastructure of the construction in premature degradation. Some factors can cause the issues of cracking formations in a concrete matrix, including mechanical load, restrained shrinkage or thermal deformation, differential settlement, poor construction methods and faulty workmanship. Moreover, conventional concrete repairing and rehabilitation techniques are time consuming and often not effective. Hence, the accurate measurements of the concrete slump components can improve the problematic issues for the better reliabilities of the concrete for the construction. As a result, the essence of high-performance concrete (HPC) is emphasized on such characteristics as high strength, high workability with good consistency, dimensional stability and durability [1]. In addition to the three basic ingredients in conventional concrete, i.e., Portland cement, fine and coarse aggregates, and water, the making of HPC needs to incorporate supplementary cementitious materials, such as fly ash and blast furnace slag, and chemical admixture, such as superplasticizer [2]. The use of fly ash and blast furnace slag, as well as other replacement materials, plays an important role in contributing to a better workability [1]. In other words, the number of properties to be adjusted has also increased results in modeling workability behavior for the concrete containing these materials is inherently more difficult than for the concrete without them. There are several studies regarding the modeling of strength of HPC, however, it is more difficult to estimate the slump and slump flow of concrete with these complex materials described above. The traditional approach used in modeling the effects of these performances of concrete starts with an assumed form of analytical equation and is followed by a regression analysis using experimental data to determine unknown coefficients in the equation [3]. However, the prediction ability of regression analyses may be limited for highly non-linear problems [4]. On the other hand, the research about the efficient data management is getting more focused and important in these days. Simultaneously, data mining techniques to deal with the reduced data without any significant meaning instead of using the original data are developed more and more. Among data mining techniques, the most frequently used techniques are applying the various multivariate analysis techniques with the extracted features by reducing the dimensionalities. In this paper, factor analysis and principal component analysis, and subtractive clustering analysis are used with maximum likelihood estimation and varimax rotation.

2 Literature review

In general, concrete consists of a mixture of paste and aggregates, or rocks. In order to make a good concrete mix, aggregates with clean, hard, strong particles free of absorbed chemicals or
coatings of clay and other fine materials that could cause the deterioration of concrete, are required. Fundamentally, the concrete consists of a mixture of aggregates and paste. First, the aggregates are comprised of inert granular materials like sand, gravel, or crushed stone mixed with water and cement. Aggregates also strongly influence concrete’s mixed and hardened properties, mixture proportions, and economy. Secondly, the paste, which is a mixture of cement and water, hardens and increases strength to form concrete through the hydration between the cement and water. The quality of the paste determines the character of the concrete as well such as increasing the strength of concrete by aging the hardening of concrete [5]. In addition, the compressive strength of concrete can be affected by the mixture ratio and curing conditions and methods of mixtures along with transporting, placing and testing the concrete. For the modernized construction using concrete, the prediction of concrete strength is the key point for the engineering judgement since it is important to know the conditions of the construction about removing concrete form, reshoring to slab, project scheduling and quality control, and the application of post tensioning for the structural engineers. Predicting the concrete strength [6, 7] has been researched for many decades based upon the concept of maturity of concrete [8, 9]. The maturity of concrete can be initially defined as “the rate of hardening at any moment is directly proportional to the amount by which the curing temperature exceeds the [datum] temperature” by McIntosh in 1949 [17]. Based on McIntosh’s concept, the same maturity of the same mix for concrete can approximately bring the same strength. In Best Practice Guide [18], the comparison of early age strength assessment of concrete has been presented by four test methods including Lok-tests, Capo tests, maturity measurement, Cube testing with air-cured cubes and temperature matched cubes, and Limpet pull-off test, along with the advantages and disadvantages.

2.1 Factor Analysis (FA)

Factor analysis [15] is a method for explaining the structure of data by explaining the correlations between variables. Factor analysis summarizes data into a few dimensions by condensing a large number of variables into a smaller set of latent variables or factors without losing any significance of the given data. Since factor analysis is a statistical procedure to identify interrelationships that exist among a large number of variables, factor analysis identifies how suites of variables are related. Factor analysis can be used for exploratory or confirmatory purposes. As an exploratory procedure, factor analysis is used to search for a possible underlying structure in the variables. In confirmatory research, the researcher evaluates how similar the actual structure of the data, as indicated by factor analysis, is to the expected structure. The major difference between exploratory and confirmatory factor analysis is that researcher has formulated hypotheses about the underlying structure of the variables when using factor analysis for confirmatory purposes. As an exploratory tool, factor analysis doesn’t have many statistical assumptions. The only real assumption is presence of relatedness between the variables as represented by the correlation coefficient. If there are no correlations, then there is no underlying structure. There are five basic factor analysis steps such as data collection and generation of the correlation matrix, partition of variance into common and unique components, extraction of initial factor solution, rotation and interpretation, and construction of scales or factor scores to use in further analyses. In addition, FA applies some rotational transformation based upon how each variable lies somewhere in the plane formed by the factors. The factor loadings, which represent the correlation between the factor and the variable, can also be thought of as the variable's coordinates on this plane. In un-rotated factor solution the Factor “axes” may not line up very well with the pattern of variables and the loadings may show no clear pattern. Factor axes can be rotated to more closely correspond to the variables and therefore become more meaningful. Relative relationships between variables are preserved. The rotation can be either orthogonal or oblique.

2.2 Principal Component Analysis (PCA)

Principal components analysis [16] is a procedure for identifying a smaller number of uncorrelated variables, called "principal components", from a large set of data. The goal of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components without losing any significance of the given data. Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. Hence, principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The following characteristics explain the PCA. First, the first principal component accounts for as much of the variability in the data as possible, and each successive component accounts for as much of the remaining variability as possible. Second,
PCA reduces attribute space from a larger number of variables to a smaller number of components and as such is a “non-dependent” procedure if it does not assume a dependent variable is specified. Third, PCA is a dimensionality reduction or data compression method. Since the goal is dimension reduction and there is no guarantee that the dimensions are interpretable, in order to select a subset of variables from a larger set based on the original variables, the highest correlations with the principal components need to be considered.

2.3 Subtractive Clustering Analysis [14]

Yager and Filev [13] in 1994 proposed a clustering method called mountain clustering, for estimating the number and initial location of cluster centers depending on the grid resolution with gridding the data space and computing a potential value for each grid point based on its distances to the actual data points. Chiu [14] suggested an extension of mountain clustering referred to as subtractive clustering, in which each data point is considered as a potential cluster centroid rather than the grid point. With Subtractive clustering method, the applied data points instead of the grid points, are evaluated independently from the dimensional problem.

The following steps summarize the Subtractive clustering method.

Step 1. Decide the measure, $M_i$, of data points, $x_i$, using

$$M_i = \sum e^{-\frac{d^2}{2r_a^2}}$$

with $\alpha = \frac{4}{r_a^2}$ where $r_a$ is a positive constant.

Step 2. Find the first center point using the highest value from Step 1.

Step 3. Recalculate the measure, $M_i$, for all data points using

$$M = M_i - M_i^* e^{-\frac{d^2}{2r_b^2}}$$

with $\beta = \frac{4}{r_b^2}$, where $r_b$ is a positive constant.

Step 4. Find the next center point using the highest value from $M_i^* < \varepsilon M_i^*$, where $\varepsilon$ is a small fraction and the value of $\varepsilon$ may generate many cluster centers if it is too small.

Step 5. Repeat this procedure until the $k^{th}$ cluster center, that satisfies the condition $M_i^* < \varepsilon M_i^*$ with the above criteria for $\varepsilon$, is calculated.

2.4 Maximum Likelihood Estimation (MLE) [11]

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters. For example, one may be interested in calculating a characteristic measurement which is unable to measure its characteristic measurement of every single object in a targeted group due to other constraints like cost or time based upon normal distribution with unknown mean and variance. At that time, the mean and variance can be estimated with MLE while only knowing the information of some sample of the overall population. In other words, the mean and variance as parameters are calculated and particular parametric values that make the observed results the most probable given the model can be identified by MLE.

2.5 Varimax Rotation [12]

The varimax rotation procedure was first proposed by Kaiser in 1958. The procedure is to find an orthonormal rotation matrix $T$ by multiplying the given number of points and the number of dimensions configuration $A$. Then, the sum of variances of the columns of $B \times B$, is a maximum, where $B = AT$. A direct solution for the optimal $T$ is not available, except for the case when the number of dimensions equals two. Kaiser suggested an iterative algorithm based on planar rotations, i.e., alternate rotations of all pairs of columns of $A$.

3 Concrete slump test data [10]

Concrete is a highly complex material. The slump flow of concrete is not only determined by the water content, but that is also influenced by other concrete ingredients. The data set includes 103 data points. There are 7 input variables, and 3 output variables in the data set. The initial data set included 78 data. After several years, 25 new data points were added to form 103 data. Seven input variables are cement, slag, fly ash, water, SP, coarse aggregate and fine aggregate. The output variable is 28-day Compressive Strength in the unit of mega pascal.

4 Applied Neurofuzzy Systems

To evaluate the acquired results, the neurofuzzy systems are used. Fig. 1 shows a neurofuzzy system...
using seven inputs and one output with the original data to evaluate the concrete compressive strength.

Fig. 1 Neurofuzzy Inference System of Compressive Strength for High Performance Concrete

Fig. 2 Neurofuzzy Inference System with membership functions of Input Variables

Fig. 3 Rulebase System for Reduced Components

Fig. 4 ANFIS Model Structure of Reduced Components

5 Analyses and results

The proposed work has been analyzed by comparing the results applying factor analysis, principal component analysis, and subtractive clustering analysis with maximum likelihood estimation and varimax rotation using the neurofuzzy systems. In Fig. 5, to determine the reduced dimensionality through a proposed procedure, the number of the reduced components or factors is determined by the accumulation of the covariance and the significant eigenvalues for the system when the eigenvalues are plotted versus each factor or component extracted by the applied multivariate analyses. As an example of using the principal component analysis among the other compared techniques, from Fig. 5, the first three or four newly extracted factors are relatively significant to implement the deployed data.
Hence, in the evaluation, three or four extracted factors by principal component analysis are used for the predicted estimation of concrete compressive strength. As a result, three and four reduced attributes are determined and used for applying the proposed techniques. Similarly, applying factor analysis, three and four newly extracted components are determined and used for evaluating the prediction of concrete compressive strength.

TABLE 1 Four Reduced Dimension Analysis

<table>
<thead>
<tr>
<th></th>
<th>corr</th>
<th>trms</th>
<th>stdev</th>
<th>mad</th>
<th>ewi</th>
<th>err</th>
</tr>
</thead>
<tbody>
<tr>
<td>org</td>
<td>0.22</td>
<td>878.2</td>
<td>84.99</td>
<td>853.1</td>
<td>1817</td>
<td>33.65</td>
</tr>
<tr>
<td>fa</td>
<td>0.55</td>
<td>4.20</td>
<td>6.69</td>
<td>4.084</td>
<td>15.4</td>
<td>3.657</td>
</tr>
<tr>
<td>favar</td>
<td>0.45</td>
<td>8.78</td>
<td>13.97</td>
<td>8.532</td>
<td>31.8</td>
<td>7.850</td>
</tr>
<tr>
<td>pcacr</td>
<td>0.41</td>
<td>8.97</td>
<td>5.91</td>
<td>4.443</td>
<td>17.8</td>
<td>4.000</td>
</tr>
<tr>
<td>pcacv</td>
<td>0.64</td>
<td>4.56</td>
<td>3.63</td>
<td>4.43</td>
<td>13.0</td>
<td>4.145</td>
</tr>
<tr>
<td>org Sub</td>
<td>0.13</td>
<td>875.6</td>
<td>85.82</td>
<td>850.6</td>
<td>1812</td>
<td>33.54</td>
</tr>
<tr>
<td>faSub</td>
<td>0.72</td>
<td>3.03</td>
<td>4.83</td>
<td>2.939</td>
<td>11.1</td>
<td>2.649</td>
</tr>
<tr>
<td>pca Sub</td>
<td>0.78</td>
<td>2.78</td>
<td>3.91</td>
<td>2.702</td>
<td>9.62</td>
<td>2.442</td>
</tr>
<tr>
<td>faml</td>
<td>0.86</td>
<td>2.96</td>
<td>2.60</td>
<td>2.876</td>
<td>8.57</td>
<td>2.822</td>
</tr>
</tbody>
</table>

Note: The employed statistical categories are Correlation (corr), Total root mean square (trms), Standard deviation (stdev), Mean average distance (mad) and Equally weighted index (ewi). The deployed neurofuzzy systems are developed by the original data (org), the original data with applying subtractive cluster analysis (orgSub), applying factor analysis using covariance (favar), applying factor analysis using maximum likelihood (faml), applying factor analysis with subtractive clustering analysis (faSub), applying principal component analysis using correlation (pcacr), applying principal component analysis using covariance (pcacv), and applying principal component analysis with subtractive clustering analysis (pcaSub).

From TABLE 1, the best performance using four reduced dimension of the data set is the prediction of concrete compressive strength applying principal component analysis with post subtracting clustering.

TABLE 2 Three Reduced Dimension Analysis

<table>
<thead>
<tr>
<th></th>
<th>corr</th>
<th>trms</th>
<th>stdev</th>
<th>mad</th>
<th>ewi</th>
<th>err</th>
</tr>
</thead>
<tbody>
<tr>
<td>org</td>
<td>0.22</td>
<td>878.2</td>
<td>84.99</td>
<td>853.1</td>
<td>1817</td>
<td>33.65</td>
</tr>
<tr>
<td>fa</td>
<td>0.55</td>
<td>4.20</td>
<td>6.69</td>
<td>4.084</td>
<td>15.4</td>
<td>3.657</td>
</tr>
<tr>
<td>favar</td>
<td>0.45</td>
<td>8.78</td>
<td>13.97</td>
<td>8.532</td>
<td>31.8</td>
<td>7.850</td>
</tr>
<tr>
<td>pcacr</td>
<td>0.41</td>
<td>8.97</td>
<td>5.91</td>
<td>4.443</td>
<td>17.8</td>
<td>4.000</td>
</tr>
<tr>
<td>pcacv</td>
<td>0.64</td>
<td>4.56</td>
<td>3.63</td>
<td>4.43</td>
<td>13.0</td>
<td>4.145</td>
</tr>
<tr>
<td>org Sub</td>
<td>0.13</td>
<td>875.6</td>
<td>85.82</td>
<td>850.6</td>
<td>1812</td>
<td>33.54</td>
</tr>
<tr>
<td>faSub</td>
<td>0.72</td>
<td>3.03</td>
<td>4.83</td>
<td>2.939</td>
<td>11.1</td>
<td>2.649</td>
</tr>
<tr>
<td>pca Sub</td>
<td>0.78</td>
<td>2.78</td>
<td>3.91</td>
<td>2.702</td>
<td>9.62</td>
<td>2.442</td>
</tr>
<tr>
<td>faml</td>
<td>0.86</td>
<td>2.96</td>
<td>2.60</td>
<td>2.876</td>
<td>8.57</td>
<td>2.822</td>
</tr>
</tbody>
</table>

For applying three reduced dimension analysis, the prediction using principal component analysis with post subtracting clustering shows the best performance among the other techniques.

6 Conclusion

As a conclusion, two analyses applying different number of reduced factors/components are compared by the estimation through the neurofuzzy systems with implemented by the reduced data set as well as the original data set to predict the concrete compressive strength. In overall, from both reduction analyses, the technique using principal component analysis with the post subtractive clustering analysis shows the best performance. However, due to the data dependency, the unique results from both cases are not presented even though the reduced cases show the better results than the original data set cases. Therefore, for the future study, more various data need to be explored to find out the better prediction of the concrete compressive strength with the various dimensionality reduction methods.

ACKNOWLEDGEMENT

Concrete Slump Test Data Set [1, 10] from UCI Machine Learning Repository are used from the following website, https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test

7 References


