Selecting an Appropriate Data Mining Algorithm to Model the Compressive Strength of High Performance Concrete

R. Uday kiran; Masaru Kitsuregawa

Institute of Industrial Science,

The University of Tokyo, Tokyo, Japan.

uday_rage@tkl.iis.u-tokyo.ac.jp; kitsure @tkl.iis.u-tokyo.ac.jp

M. Venu

Birla Institute of Technology and Science, Pilani

Hyderabad campus, Hyderabad, Andhra Pradesh, India.

venu.bits@gmail.com

Abstract

An open problem in the research of building materials is the selection of an appropriate algorithm to model compressive strength of High Performance Concrete (HPC) effectively. It is because there is no data mining algorithm that outperforms others for all performance measures. Every measure is equally important as each of them highlights different features of the data and of model behavior. The field Multi-Criterion Decision Making (MCDM) involves structuring and solving the decision problems involving multiple criteria where there does not exist a unique optimal solution, and it is necessary to use the preferences of a decision maker to differentiate between solutions. This paper makes an effort to address the problem of algorithm selection by posing it as a MCDM problem, and suggest a solution using Analytical Hierarchy Process. Experimental results show that the proposed technique facilitates the user to select an appropriate algorithm to model compressive strength of HPC.

1. Introduction

1.1. Background

High Performance Concrete (HPC) is an important element in modern infrastructure development. It is a highly complex mixture constituting of concrete and supplementary cementing ingredients. The concrete ingredients are cement, fine and coarse aggregates, and water. The supplementary cementing ingredients are the materials used to make concrete mixtures more economical, reduce permeability, increase strength, or influence other concrete properties. Examples include fly ash, blast furnace slag and superplasticizer. The compressive strength of concrete (or concrete strength) plays a key role in mix design and quality control. Thus, modeling of concrete strength is an important step in building materials.

The Abram's water-to-cement ratio has been widely used in the past to model concrete strength [1]. This measure predicts the concrete strength using the proportions of water and cement ingredients, and disregards the proportions of other ingredients. Several studies independently have shown that concrete strength development is influenced not only by water-to-cement ratio, but also by the proportions of other ingredients. Since then the modeling of concrete strength has become a difficult task in building materials.

Over the past few decades, artificial neural networks (ANN) have been extensively used by the civil engineering researchers to model behavior of different materials [2]. It is because ANN can capture highly non-linear and complex relation between input/output variables in an application without any prior knowledge about the nature of interaction. Most of the ANN-based concrete strength prediction models are based only on the tests of concrete without considering supplementary cementing ingredients [3, 4, 5]. Yeh proposed an ANN model to predict the concrete strength effectively using both concrete and supplementary cementing ingredients [6, 7]. The influence of ingredients in modeling of concrete strength still remains unclear due to the "black box" nature of ANN. Venu et al. [8] have applied feature selection techniques to understand the influence of ingredients, and proposed a robust ANN model to predict concrete strength effectively.

1.2. Motivation

Since the introduction of data mining in [9], numerous predictive data mining algorithms have been proposed in the literature to model various real-world applications. Selecting an appropriate algorithm is an important problem in data mining. It is because there exists no universally acceptable best algorithm that outperforms others for all datasets and performance measures. In the literature, researchers tried to confront the problem by combining different individual algorithms using meta-learning techniques, such as bagging and boosting [10, 11]. Recently, Chou et al. [12] have investigated the performance of different individual and meta-learning algorithms to model the concrete strength, and showed that even the meta-learning algorithms have not outperformed the individual algorithms on all performance measures. Thus, selecting an appropriate algorithm to model concrete strength still remains an open problem in building materials research.

1.3. Contribution of this Paper

This paper proposes a technique of selecting an appropriate algorithm to model concrete strength effectively. The technique involves formulating the problem of selecting a right algorithm as a multi-criteria decision making problem and suggesting the solution using the Analytical Hierarchy Process (AHP). Please note that the proposed technique can be extended to select an appropriate algorithm to model any real-world application, where there exists no algorithm that outperforms others over all the performance measures.

1.4. Paper Organization

The rest of the paper is organized as follows. Section 2 discusses the previous works on modeling of concrete strength. Section 3 introduces Analytical Hierarchy Process and describes the proposed technique of selecting an appropriate algorithm to model concrete strength. Experimental results are presented in Section 4. The paper concludes with future research directions.

2. Related Work

Since the initial application of ANN in [14], researchers have made efforts to model different aspects of building materials using ANN [2]. Most of these works used only concrete ingredients to model the strength.

A Fuzzy-Neuro model was used by Nataraj *et al.* to predict the compressive strength of concrete designed as suggested in IS10262-2003 and IS456-2000 [15]. Gupta [16] worked on the modeling of concrete strength using Support Vector Machines. Radial basis function (RBF) and polynomial kernels are used with support vector machines. Results from the experiments suggest that support vector machine based modeling approach can effectively be used in predicting the compressive strength of high performance concrete.

Rajiv Gupta *et al.* [17] presented a neural-fuzzy inference system for predicting the compressive strength of HPC. The system parameters included concrete mix-design, specimen size and shape, curing technique and period, and the environmental conditions, such as maximum temperature, relative humidity, and wind velocity. A data driven rule-based expert system was developed to overcome the bottlenecks in knowledge acquisition. Although this reduced predictive accuracy, the system enabled the easy update of the knowledge base at any stage to enhance the neural-expert interface.

Yen [6, 7] made an effort to model the strength using both concrete and supplementary cementing ingredients. In [6], it was shown that ANN outperforms the regression. A software toolkit HPC2N (High Performance Concrete Design Package Using Neural Network and Nonlinear Programming), which uses ANN and linear programming techniques has been proposed in [7] to find optimal mixing proportions of various ingredients for desired concrete strength. The input attributes used in model preparation are *cement, water, fine aggregate, coarse aggregate, blast furnace slag, superplasticizer, fly ash* and *age of concrete*. The influence of ingredients in modeling of concrete strength still remains unclear due to the "black box" nature of ANN. Venu et al. [8] have applied feature selection techniques to understand the influence of ingredients, and shown that the concrete ingredients, *fine* and *coarse aggregates*, are irrelevant or redundant with respect to modeling of concrete strength, and therefore, can be neglected while modeling the concrete strength using ANN.

Chou *et al.* [12] have investigated the performance of different data mining algorithms on the prediction of concrete strength, and reported that there exists no algorithm that outperforms others for all parameter measures. The authors have not discussed any mechanism for selecting an appropriate algorithm to handle such cases. Thus, selecting an appropriate algorithm to model concrete strength still remains an open problem in the research area of building materials. This paper tries to address the problem using AHP.

3. Analytical Hierarchy Process

Multiple-Criteria Decision Making (MCDM) is an important area in operations research. It is concerned with structuring and solving decision and planning problems involving multiple criteria where there does not exist a unique optimal solution and it is necessary to use decision maker' preferences to differentiate between solutions.

AHP [13] is an important method in MCDM. It consists of three stages of problem-solving: decomposition, comparative judgments, and synthesis of priority. The decomposition stage aims at the construction of a hierarchical network to represent a decision problem, with the top level representing overall objectives and the lower levels representing criteria, sub-criteria, and alternatives (see Figure 1(a)). With comparative judgments, users are requested to set up a comparison matrix at each hierarchy by comparing pairs of criteria or sub-criteria. Table 1 shows the fundamental scale suggested by T. L. Satty for the pair-wise comparisons. Finally, in the synthesis of priority stage, each comparison matrix is then solved by an eigenvector method for determining the criteria importance and alternative performance.

Table 1. Fundamental scales for pair-wise comparisons

Verbal Scale	Numerical Values
Equally important, likely or preferred	1
Moderately more important, likely or preferred	3
Strongly more important, likely or preferred	5
Very strongly more important, likely or preferred	7
Extremely more important, likely or preferred	9
Intermediate values to reflect compromise	2,4,6,8
Reciprocals for inverse comparison	Reciprocals

The hierarchical structure used in this paper for selecting an algorithm model concrete strength is shown in Figure 1(b). In the hierarchical structure, the goal was the select an appropriate algorithm, the criterions were the performance measures and the alternatives were the predictive data mining algorithms.

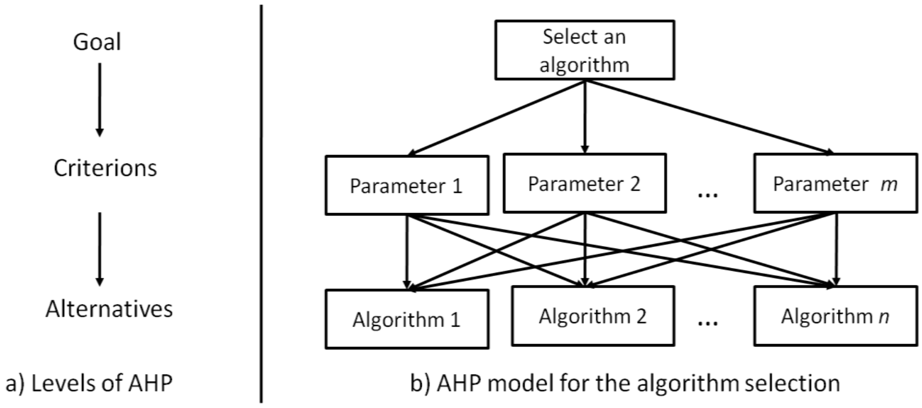

Figure 1: Hierarchical model for the algorithmic selection.

The following section describes the methodology of selecting an appropriate algorithm to model concrete strength using the proposed hierarchical structure.

4. Experimental Results

4.1. Dataset Description

The experimental data was obtained from the University of California, Irvine (UCI) repository of data [20]. It contains 1030 samples of HPC and published by Yen [6]. Table 2 shows the experimental dataset of nine HPC attributes used in this study. The first eight attributes are input attributes and the last attribute is a prediction attribute.

Table 2. Details of the data set.

S. No.	Attributes	Unit	Range of values	Mean	Standard Deviation	Distant values
1	Cement	kg/m^3	[102, 540]	281.1	104	278
2	Blast furnace slag	kg/m^3	[11, 359.4]	107.3	61.9	185
3	Fly ash	kg/m^3	[24.5, 200.1]	83.9	39.98	156
4	Water	kg/m^3	[121.8, 247]	181.567	24.4	195
5	Superplasticizer	kg/m^3	[1.7, 32.2]	8.5	4	111
6	Coarse Aggregate	kg/m^3	[801, 1145]	972.9	77.75	284
7	Fine Aggregate	kg/m^3	[594, 992.6]	773.6	80.17	302
8	Age	Day	[1, 365]	45.66	63.17	14
9	Compressive Strength	MPa	[2.3, 82.6]	35.8	16.7	845

4.2. A Comparative Evaluation of Data Mining Algorithms

Chou et al. [12] have evaluated the performance of different data mining algorithms against the various parameters using the stratified *k*-fold cross-validation with *k* set to 10. The algorithms used are Artificial Neural Networks (ANN), Multiple Regression (MR), Support Vector Machine (SVM), Multiple Additive Regression Trees (MART), Bagging Regression Trees (BRT). The parameters are Correlation Coefficient (R), Coefficient of Determination (R^2), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Table 3 shows the mean performance of each algorithm for each parameter. (To model a problem with AHP, maximum value of a parameter must be preferred the user. However, minimal value is preferred by the user for the parameters, *RMSE* and *MAPE*. Therefore, their reciprocals (i.e., $RMSE^{-1}$ and $MAPE^{-1}$) have to be considered in AHP.)

Table 3: Performance comparison of the predictive data mining algorithms. The values in bold represents the best algorithm to model concrete strength for the corresponding measure.

Algorithm	Average R	Average R^2	Average $RMSE$	(Average $RMSE)^{-1}$	Average $MAPE$	(Average $MAPE)^{-1}$
ANN	**0.95**	0.90	5.0	0.2	**10.9**	0.1
MR	0.78	0.61	10.4	0.1	31.65	0.03
SVM	0.94	0.88	5.6	0.18	12.77	0.08
MART	**0.95**	**0.91**	**4.9**	0.2	13.89	0.07
BRT	0.94	0.89	5.6	0.18	14.18	0.07

It can be observed that there exists no algorithm that outperforms others over all parameter measures. Chou *et al.* [12] has suggested the usage of MART as it outperformed others (especially, its competitor ANN) in more number of parameter measures. However, such an approach of selecting an algorithm may not be a rational method for taking a decision. It is because, although ANN has fallen back over MART by a very smaller margin for the parameters R^2 and RMSE, MART has fallen back over ANN by a very larger margin for the parameter MAPE. Thus, selecting an appropriate algorithm to model concrete strength has become a very difficult task. We now discuss the algorithm selection using AHP.

4.3. Algorithm Selection using AHP

After constructing hierarchy as shown in Figure 1(b), a pair-wise comparison matrix for all parameters in a level is constructed by quantifying the domain experts preferences using a nine-point scale shown in Table 1. Table 5 shows the comparative preference scores provided by the domain expert to model concrete strength. The pair-wise comparison table generated using Table 5 is shown in Table 6. Table 7 shows the eign vector (or weights) derived from the pair-wise comparison matrix by normalizing the summation of rows values. For simplicity, we call the column matrix *A*.

Table 5: Comparative preferences of the domain expert for the measures.

$MAPE^{-1}$ vs. R	4:1
$RMSE^{-1}$ vs. R	2:1
R^2 vs. R	2:1
$MAPE^{-1}$ vs. $RMSE^{-1}$	1:1
$MAPE^{-1}$ vs. R^2	2:1
$RMSE^{-1}$ vs. R^2	2:1

Table 6: Pair-wise comparison of the performance measures. To minimize the user's burn, the closure property has been used to derive comparative preferences for the measures that were not provided by the user.

	R	**R2**	**$RMSE^{-1}$**	**$MAPE^{-1}$**
R	1	2	4	4
R2	0.5	1	2	2
$RMSE^{-1}$	0.25	0.5	1	1
$MAPE^{-1}$	0.25	0.5	1	1

Table 7: Weights (or Eigen vector) of the performance measures in percentages.

Measure	**Weight (%)**
R	50
R^2	25
$RMSE^{-1}$	12.5
$MAPE^{-1}$	12.5

After determining the priorities of parameters, an eign vector for each parameter is generated by normalizing the performances of all algorithms with respect the corresponding measure (see Table 10). For example, the summation of performance values of all algorithms for the parameter *R* is 4.56 (= 0.95+0.78+0.94+0.95+0.94). The normalized performance of an algorithm , say ANN, for the parameter *R* in percentage is 20.83 (=0.95*100/4.56). Similarly, for the algorithms, MR, SVM, MART and BRT, the values will be 17.1, 20.61, 20.83, 20.61, respectively (see the second row of Table 10). Let this matrix be named as *B*.

Table 10: Comparison of the algorithms against the performance measures.

	ANN	SVM	MR	MART	BRT
R	20.89	20.62	17.1	20.91	20.48
R2	21.61	21.05	14.53	21.65	21.16
$RMSE^{-1}$	23.26	20.93	11.63	23.26	20.93
$MAPE^{-1}$	28.57	22.86	8.57	20	20

Aggregate the relative weights up the hierarchy to obtain a composite weight which represents the relative importance of each alternative according to the decision-maker's assessment. In other words, perform the multiplication with matrices *A* and *B* (i.e., A x B) to derive a column matrix, say *C*, which represents the weighted value of each algorithm for user-given parameters' priorities. Table 11 shows the weighted value and rank of each algorithm. The results show that ANN is an appropriate algorithm to model concrete strength for the user-given preferences. From Table 3, it can be observed although ANN was the best for some of the parameters, it was falling back by a very little margin only.

Table 11: Ranking of the algorithms.

Algorithms	Weighted value	Rank
ANN	24.68	1
SVM	21.64	2
MART	21.46	3
BRT	20.57	4
MR	11.65	5

Evangelos Triantaphyllou and Stuart H. Mann [19], have reviewed applications of MCDM methods on many engineering disciplines and reported that MCDM methods, such as AHP, should be used as decision support tools and not as the means for deriving the final answer. The reason is that finding the truly best solution to a MCDM problem may never be humanly possible.

6. Conclusions and Future Work

This paper has made an effort to address the open problem of selecting an appropriate algorithm to model concrete strength using AHP. A novel hierarchical structure has been introduced in this paper. Empirical results on a real-world dataset show that the proposed technique suggests the user (or domain expert) an appropriate algorithm that satisfies the given preference criteria of parameters.

Developing a generalized framework involving the comparison of various data mining algorithms on a civil engineering dataset, selecting an algorithm and building a robust prediction model by applying feature selection techniques on the selected algorithm is an important future work of the paper. Another future work will be extending the proposed technique to different civil engineering applications (or datasets).

7. References

[1] Aitcin, Pierre-Claude and Neville, Adam (1993), *High-performance concrete dymystified*, *Concrete International,* Vol. 15, no. 1, pp. 21-26.

[2] H. Adeli (2001), *Neural Networks in Civil Engineering: 1989 - 2000, Computer-Aided Civil and Infrastructure Engineering 16,* pp. 126-142.

[3] Jong-In Kim, Doo Kie Kim, Maria Q. Feng, Frank Yazdani (2004), *Application of Neural Networks for Estimation of Concrete Strength*, Journal of Materials in Civil Engineering, Vol. 16, Issue 3, pp. 3-16.

[4] Jun Peng, Zongjin Li and Baoguo Ma (2002), *Neural Network Analysis of Chloride Diffusion in Concrete, Journal of Materials in Civil Engineering,* Vol. 14, Issue 4, pp. 327-333.

[5] J. A. Stegemann and N. R. Buenfeld (2004), Mining of Existing Data for Cement-Solidified Wastes Using Neural Networks, Journal of Environmental Engineering, Vol. 130, Issue 5, pp. 508-516.

[6] 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.

[7] I-Cheng Yeh (2006), *Analysis of Strength of Concrete using Design and Experiments of Neural Networks, Journal of Materials in Civil Engineering,* Vol. 18, No. 4, pp. 597-604.

[8] M. Venu, R. Uday kiran and R. Kiranmai (2011), A Robust Neural Network Classifier to Model the Compressive Strength of High Performance Concrete using Feature Subset Selection, ACM Compute, India.

[9] Rakesh Agrawal and Ramakrishnan Srikanth (1994), *Fast Algorithms for Mining Association Rules, Very Large Database Conference,* pp. 487-499, Chile.

[10] D. Fan, P. Chan and S. Stolfo (1996), *A Comparative Evaluation of Combiner and Stacked Generalization, AAAI 1996 Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, Association for the Advancement of Artificial Intelligence*, pp. 40-46.

[11] J.H. Freedman (2000), *Greedy Function Approximation: A gradient Boosting Machine, Annals of Statistics*, Vol. 25, Issue 5, pp. 1189-1232.

[12] Jui-Sheng Chou, Chien-Kuo Chiu, Mahmoud Farfoura and Ismali Al-Taharwa (2011), Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques, Journal of Computing in Civil Engineering, Vol. 25, Issue 3, pp. 242-253.

[13] Thomas L. Saaty (1980), *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*, McGraw-Hill.

[14] H. Adeli and C. Yeh (1989), Perceptron Learning in Engineering Design, Microcomputers in Civil Engineering, pp. 247-256, USA.

[15] M. C. Nataraja, M. A. Jayaram and C. N. Ravikumar (2006), *A Fuzzy-Neuro Model for Normal Concrete Mix Design, Engineering Letters,* Vol. 12, Number 2, pp. 98-107.

[16] Gupta (2011), *Support Vector Machines based Modeling of Concrete Strength, World of Academy of Science,* Vol. 36, pp. 305-311.

[17] M. A. Rajiv Gupta, Kewalramani and A. Goel (2006), *Prediction of Concrete Strength using Neural-Expert System, Journal of Materials in Civil Engineering,* Vol. 18. Issue 3, pp. 462-467.

[18] Meltem, Özturan, Birgul Kutlu and Turan Özturan (2008), *Comparison of Concrete Strength Prediction Techniques With Artificial Neural Network Approach, Building Research Journal*, Volume 56, pp. 23-36.

[19] E. Triantaphyllou and S.H. Mann (1995), *Using the Analytic Hierarchy Process For Decision Making in Engineering Applications: Some Challenges, Journal of Industrial Engineering: Applications and Practice,* Volume 2, Issue 1, pp. 35-44.