Comparison of Machine Learning Techniques for Developing Performance Prediction Models

Nima Kargah-Ostadi, Ph.D.[1]

[1]Fugro Roadware, Inc.; 8613 Cross Park Drive, Austin, TX 78754; PH (512) 977-1883; FAX (512) 973-9565; email: nkargah-ostadi@fugro.com

ABSTRACT

A central component of every decision support system is a performance prediction model. Often, empirical data is used to describe a performance measure as a function of factors affecting performance. In this context, hosts of machine learning techniques have demonstrated function approximation capabilities beyond regression methods. This study is aimed at comprehensive comparison of nine machine learning algorithms for developing a pavement performance model. Instrumental in pavement management systems, this extensive evaluation reconciles computational intelligence with systems engineering. Three principles of maximum likelihood, consistency, and parsimony are considered in algorithm benchmarking and model development. Along with quantitative measures of accuracy and generalization, qualitative criteria including scatter plots and sensitivity analysis are also implemented. Additionally, quantitative and qualitative aspects of model evaluation are combined through analysis of predicted performance trends. Results indicate that an algorithm performing the best in each comparison step might not achieve the same in the other. This comparison framework is recommended for identifying the appropriate machine learning paradigm to develop performance prediction models.

INTRODUCTION

A pavement management system (PMS) is a set of analytic tools developed based on systems engineering to assist roadway decision-making. PMS are decision support systems for construction, maintenance, and rehabilitation of pavement structures. Similar to every decision support system, performance prediction models are a central part of every network-level PMS to predict future performance of the pavement network, identify maintenance and rehabilitation (M&R) needs, and estimate the network conditions after the application of various treatment alternatives (Haas et al. 1994). Performance models are implemented to guide M&R project selection, together with life cycle cost analysis (LCCA) and engineering judgment.

Change in pavement surface roughness over time is one of the most important performance indicators, because it affects vehicle ride quality and dynamic loads. One of the main objectives of every road authority is to provide a comfortable ride for users, and pavement roughness is a good indicator of whether this criterion will be

fulfilled. Therefore, International Roughness Index (IRI) has been used in PMS as the major indicator of pavement functional performance (Haas et al. 1994).

Various studies have been conducted regarding pavement deterioration trends and factors affecting performance. Most of these studies have limitations such as the correlation of input variables, difficulty of variables data collection, and experimental shape of performance curves, among others (Perera and Kohn 2001). Consequently, the need for a simpler and more efficient empirical model to use in network-level PMS is still an ongoing pursuit of the pavement engineering community.

Due to the large number of variables and the complex ways in which they affect one another, use of simple statistical approaches such as linear regression is not a viable means to develop pavement roughness prediction models. In addition, the shape of pavement performance curve is not known beforehand and multiple arrangements have to be tested in order to develop a model using nonlinear regression (Von Quintus et al. 2001). Hence, studies have attempted to use computational intelligence techniques such as artificial neural networks (ANN) to develop more accurate models (Attoh-Okine 1994; Kargah-Ostadi et al. 2010).

Past studies have largely implemented feed-forward ANN to develop pavement performance models. Other promising machine learning techniques exist that could potentially be suitable for pavement performance modeling. Most of the previously developed models have been evaluated based on prediction accuracy only and a few have evaluated generalization capability. Among others, one major limitation of these studies is lack of qualitative evaluations of model performance.

In this study, data from federal highway administration (FHWA)'s long-term pavement performance (LTPP) program are extracted and preprocessed for model development. A deterministic modeling approach is adopted in which pavement IRI is considered as the dependent variable and factors affecting roughness are used as input variables. In this context, a range of machine learning techniques including ANN, radial basis function (RBF) networks, and support vector machines (SVM) are benchmarked on the current function approximation problem and compared against each other. Along with error measures of IRI estimation, generalization capability and error in predicting pavement deterioration rate are also compared among various methods on a testing database that has not been used for model development.

MACHINE LEARNING

Machine learning (ML) techniques are a sub-category of computational intelligence techniques mainly employed for deriving definitive information out of large sets of data for pattern recognition, classification, function approximation, and so forth (Jain et al. 1996). Widely used for function approximation problems, a subset of ML algorithms including ANN, RBF networks, and SVM have very similar designs involving a large number of parallel but connected simple processors. In the ANN terminology, these computing nodes are called neurons, which are organized in a number of layers and connected to each other with model parameters.

These similar networks of processors are trained using different learning paradigms to estimate model parameters based on observed data. While ANNs are inspired by biological neurons, RBF networks and SVM are based on statistical

learning theory. The ANN learning paradigm is a recursive stochastic approximation used in supervised training of multi-layer perceptrons (MLP). The RBF learning paradigm is termed hyper-surface reconstruction, which is essentially smooth curve fitting using regularized interpolation. The SVM learning paradigm is an approximate implementation of the principle of structural risk minimization adopted from the statistical learning theory (Haykin 1999). The maximum degradation (distance) of an approximating hyper-plane from the observed data is minimized.

Past studies proved the Universal Approximation Theorem for ANN, stating that an MLP having a single hidden layer is adequate to approximate any continuous nonlinear input-output relationship to any degree of accuracy. However, this does not mean that the MLP would be the most efficient one or that it would have a good generalization capability. Similar studies have proved that RBF networks and SVM are also universal approximators (Haykin 1999). Any superior performance of each technique compared to others is deemed to be problem specific. Hence, algorithms need to be benchmarked on each problem to determine the suitable architecture for each algorithm and to compare alternative algorithms.

DATA EXTRACTION AND PREPROCESSING

Data from LTPP flexible pavement experiments were used in developing the IRI models (LTPP 2013). With diverse structural factors and alternative rehabilitation treatments, these pavement sections have been constructed in a variety of climatic regions and on various subgrade soil types. Factors affecting pavement surface roughness were used as input variables to develop a model to predict the output variable, roughness. Input variables were selected based on a comprehensive literature review and LTPP data availability. Details of input variable selection and compilation of data records are available elsewhere (Kargah-Ostadi 2013).

Several preprocessing steps were taken for smoothing, outlier detection, normalization (mean removal), and de-correlation. In order to alleviate the irrational fluctuations in the time dependent performance data, a time-normalized three-point moving average scheme was used for smoothing the performance versus age curves. As the pavement deteriorates, roughness is typically increasing with time. Data records where the IRI measurement was less than its previous IRI reading (while no maintenance or rehabilitation treatment had happened) were considered outliers and were eliminated from the database. Next, the input variables were normalized to have zero mean and a standard deviation of one. The output variable was mapped to a range of [-1, +1] for consistency.

After normalization and outlier detection, principal component analysis (PCA) was used to simplify the developed models through reduced dimensionality and eliminate input variable correlations, thereby reducing over-fitting probability. Principal components (PC) are new uncorrelated variables created through a linear combination of correlated input variables (Jolliffe 1986). Through decomposition of the covariance matrix of the original correlated inputs, PCA arranges the resultant components in an order according to the portion of the total input data variance that they account for. In this study, the first ten components were selected as uncorrelated input components used for model development, because they explain more than 95

percent of the total variance. Unlike stepwise regression, PCA allows implementation of all input variables in model development.

Data records for one section from each climatic region were selected randomly for testing of the developed models regarding roughness progression trends. The remaining 3361 data records were divided into training (70%), validation (15%), and testing (15%) datasets. The validation dataset was used in early stopping learning paradigms in order to avoid over-fitting due to noise. In other generalization learning algorithms, the validation and training datasets were jointly used for model development. For quantitative evaluation of model prediction accuracy, the testing dataset was used because it had not been utilized at the development stage.

ALGORITHM BENCHMARKING AND MODEL DEVELOPMENT

Table 1 lists the implemented learning machines. For improving generalization in neural networks, two approaches were compared: early stopping using the validation dataset, and regularization using Bayesian inference technique.

Table 1. Considered learning machines for IRI model development.

Learning Paradigm	Machine	Training Approach	Symbol
ANN	Feed-Forward MLP	Levenberg-Marquadt (LM) with Early Stopping	FF-ANN-ES
		LM with Bayesian Regularization	FF-ANN-BR
	Cascade-Forward MLP	LM with Early Stopping	CF-ANN-ES
		LM with Bayesian Regularization	CF-ANN-BR
RBF Networks	Generalized RBF Network	Adaptive Number of Centers	AG-RBF
		Fixed Number of Centers	G-RBF
	Generalized Regression Neural Network	Strict Interpolation with Normalized Gaussian Kernels	GRNN
SVM	Support Vector Regression	Polynomial Kernels	P-SVM
		Gaussian Kernels	G-SVM

In early stopping, the training patterns are divided randomly into training and validation subsets. Mean squared error (MSE) on training subset is minimized until the MSE on validation subset starts to increase. The training process is then stopped at the minimum MSE on validation data. In regularization training, a regularizing term is added to be minimized along with the MSE. This additional term typically penalizes larger weights. Therefore, this type of regularization is called weight decay and advocates smaller weights for less important network connections. One way or another, all of the considered machines attempt to balance bias (lack of fit) and variance (over-fitting due to noise) in order to improve generalization.

As most of the involved learning processes are local optimization techniques, multiple seed analyses are required for a sound comparison of architectures within each paradigm. For each learning machine, various architectures were compared according to the procedure depicted in Figure 1 so that the architecture with highest accuracy and generalization was chosen. By seeking the most successful training instance, all of the nine different learning models are benchmarked on the given

function approximation problem. Since there is a significant variance in the training results of alternative architectures for each learning paradigm, different learning machines can only be compared after such benchmarking procedure is realized.

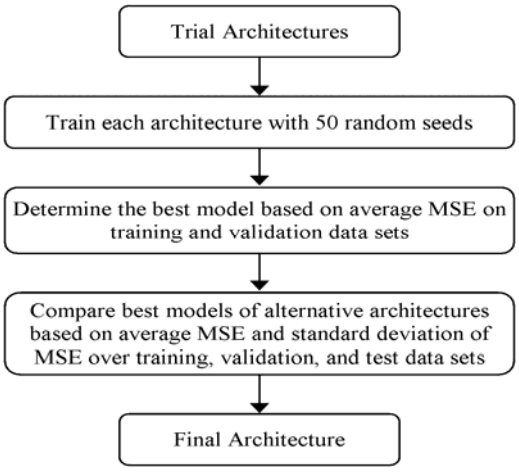

Figure 1. Benchmarking flowchart for each learning machine.

Alternative architectures for ANN were developed via changing number of layers and neurons, and the transfer functions. In RBF networks, spread of the centers (σ), the regularizing parameter (λ), and the size of the hidden layer (fraction of training patterns) were altered among architecture options. For SVM models, the architecture choices were created with varying the type of kernel function, spread of the centers of RBF kernels, insensitivity margin (ε), and the regularizing parameter (C). Details of the final selected models are available elsewhere (Kargah-Ostadi 2013).

RESULTS AND DISCUSSION

In order to provide a comprehensive comparison of various machine learning techniques, both the learning process and the developed models need to be evaluated (Figure 2). The learning processes were compared in terms of effectiveness (average MSE of the final model), efficiency (total amount of time required to evaluate alternative architectures and determine the optimum), and reliability (subtracting from 100, the coefficient of variation of MSE among 50 random trials of the algorithm).

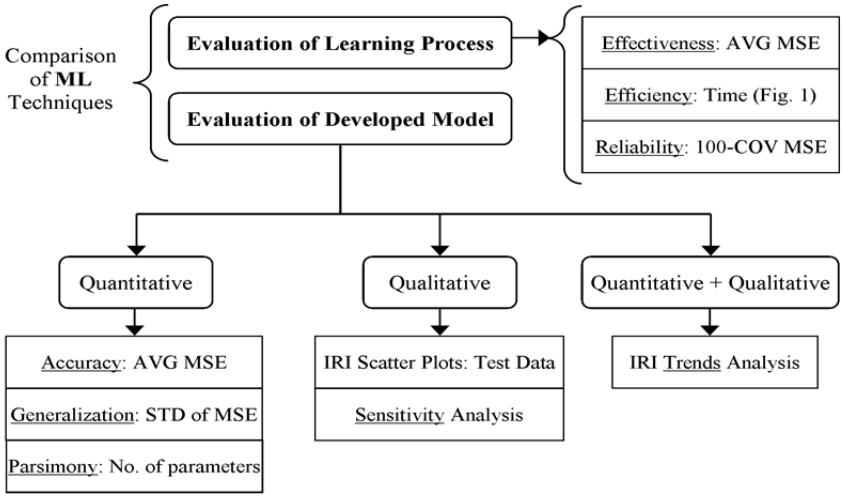

Figure 2. Framework for comprehensive comparison of various machine learning techniques in development of pavement performance models.

Compared to regularization, MacKay (1992) describes early stopping as a "patching up of a bad model." While having lower efficiency (taking 18 times the processing time) and lower reliability (80 percent), Bayesian regularization seemed to be twice more effective (achieving half the MSE) compared to early stopping for ANN. Gaussian SVM and generalized RBF networks with fixed centers (G-RBF) followed regularized ANN in terms of effectiveness. Generalized regression neural networks were the most efficient model development method, since there are few architectural parameters that can be altered and each training process takes little time to complete. Most of the learning methods have high reliability (above 90 percent).

In this study, three important principles of model development (Gupta et al. 2008) were considered: maximum likelihood (minimizing MSE on training data), consistency (generalization), and parsimony (selecting the simplest learning network, with equal error). The same principles were used in quantitative and qualitative evaluation of the final developed models (Figure 2). For quantitative evaluation, accuracy and generalization capability of the models were determined, respectively using average and standard deviation of MSE over training and test datasets. In addition, model complexity was represented as the number of parameters involved.

Figure 3 represents the quantitative comparison of the nine IRI models. FF-ANN-BR was the most accurate and had the best generalization capability. However, if model parsimony is considered, CF-ANN-BR was the best option. It is important to note that for some models, such as the GRNN, there was a significant difference between the accuracy and generalization capability.

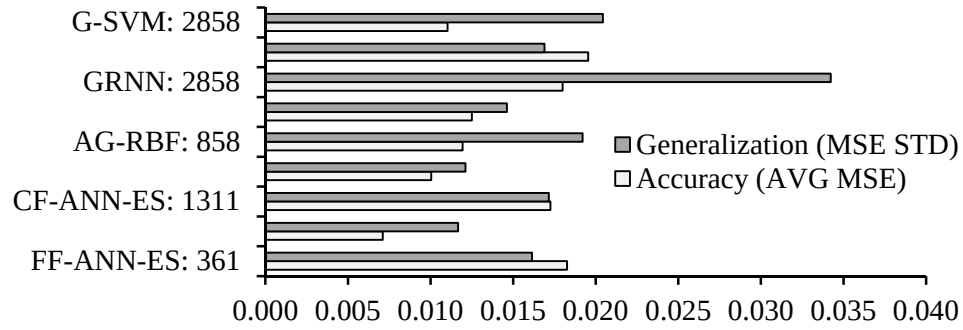

Figure 3. Quantitative comparison of accuracy, generalization, and parsimony of the final developed models; number of model parameters is indicated in front of each model name; accuracy was calculated as the average MSE over training and test datasets; generalization was measured as the standard deviation of MSE between datasets.

Four models were selected based on quantitative evaluation (Figure 3) to represent the best models for each learning paradigm: FF-ANN-BR, CF-ANN-BR, G-RBF, and G-SVM. For qualitative evaluation, scatter-plots of measured versus predicted IRI over the testing database (504 records not used in model development) were examined. The FF-ANN-BR (R^2=0.94), CF-ANN-BR (R^2=0.93), G-RBF (R^2=0.91), and G-SVM (R^2=0.90) in this order provided the best scatter plots and this agrees with the quantitative evaluation as well.

Another qualitative evaluation tool is sensitivity analysis of the models to important input factors, in order to evaluate consistency in model function based on principles of pavement engineering. It should be noted that regarding machine

learning techniques, evaluation of model form is not viable in the same format as nonlinear regression techniques. Nevertheless, the developed models are intended to support prediction rather than scientific reasoning.

All but one input variable were kept constant at database average values and one variable was increased or decreased by 25 percent to examine sensitivity of model output to each input factor. Generally, the CF-ANN-BR and G-RBF models provided a sensitivity pattern that agreed with previous studies of factors affecting flexible pavement roughness (Perera and Kohn 2001). Details of the sensitivity analysis are available elsewhere (Kargah-Ostadi 2013).

A combination of quantitative and qualitative evaluations was also carried out by checking variation of model output with time (Gupta et al. 2008). Not previously implemented, data records from one section in each of the four climatic regions were used to compare average absolute error (AAE) of observed and predicted IRI change rates (Figure 4). While the generalized RBF network (G-RBF and AG-RBF) models did not have the best quantitative capacities (Figure 3), they performed better than the other models in predicting roughness progression rates.

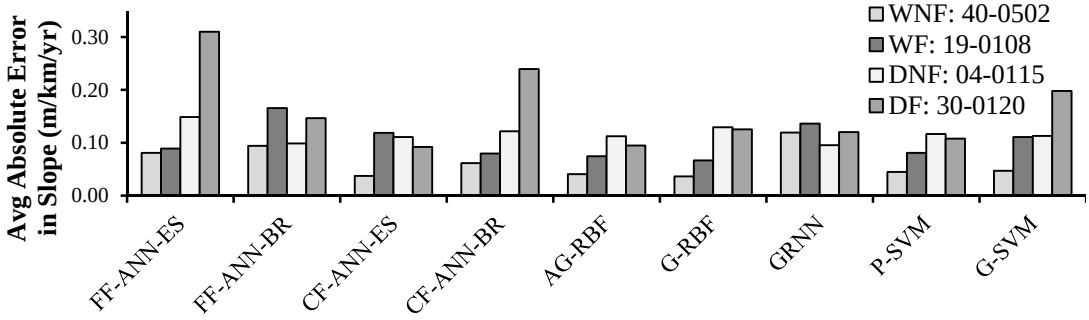

Figure 4. Comparison of AAE of the models in predicting IRI change rate in each of the four climatic regions (Wet Non-Freeze, Wet Freeze, Dry Non-Freeze, and Dry Freeze).

The overall mixture of quantitative and qualitative evaluations of the developed IRI models gave an edge to the generalized RBF network with fixed centers (G-RBF). While it had moderate accuracy and generalization capability compared to other models, it performed best in terms of sensitivity behavior and prediction of roughness progression trends. This comparison framework demonstrates that a model having better accuracy might not necessarily generalize well. At the same time, quantitative evaluation alone is not sufficient to determine the best model and other qualitative measures should also be considered.

CONCLUSIONS AND RECOMMENDATIONS

In this study, several machine learning techniques were compared in developing pavement roughness prediction models. Preprocessing steps included smoothing, outlier detection, normalization, and de-correlation via principal component analysis. Variant architectures of neural networks, radial basis function networks, and support vector machines were benchmarked to determine the optimum parameters. The different learning processes were compared in terms of effectiveness, efficiency, and reliability. While having lower efficiency and reliability, Bayesian regularization was the most effective learning algorithm for ANN.

Following the benchmarking of the algorithms on the specific performance modeling problem, the final developed models were compared through quantitative and qualitative evaluations. Quantitative evaluation revealed that having higher accuracy (lower average error on training data) does not always result in models that generalize well. The regularized ANN models had the best performance in terms of accuracy and generalization capability. Qualitative evaluations included scatter plots of predicted versus measured IRI and sensitivity analysis of the models with respect to variations in factors affecting roughness. Regularized cascade-forward ANN and generalized RBF networks had the best qualitative performance compared to others.

Combining quantitative and qualitative aspects, an analysis of predicted roughness progression rates indicated that the IRI output of the generalized RBF network most resembled measured trends. Overall, the generalized RBF network model has an edge over other models. This comparison, while comprehensive, is problem specific and should be repeated for other performance modeling efforts.

REFERENCES

Attoh-Okine, N. O. (1994). "Predicting Roughness Progression in Flexible Pavements Using Artificial Neural Networks," *Proceedings of the 3rd International Conference On Managing Pavements*, Vol. 1, San Antonio, Texas, pp. 55-62.

Gupta, H.V., Wagener, T., and Yuqiong, L. (2008). "Reconciling Theory with Observations: Elements of a Diagnostic Approach to Model Evaluation," *Hydrol Process*, Vol. 22, pp. 3802-3813.

Haas, R., Hudson, W. R., and Zaniewski, J. (1994). *Modern Pavement Management*, Krieger Publishing Company, Malabar, Florida.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. 2nd Edition, Prentice Hall, New Jersey.

Jain, A. K., Mao, J., and Mohiuddin, K. M. (1996). "Artificial Neural Networks: A Tutorial," In IEEE *Computer*, Vol. 29, No. 3, pp. 31-44.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, NewYork.

Kargah-Ostadi, N., Stoffels, S. M., and Tabatabaee, N. (2010). "Network-Level Pavement Roughness Model for Rehabilitation Recommendations." *Transportation Research Record*, No. 2155, pp. 124-133.

Kargah-Ostadi, N. (2013). *Enhancing Analytical Toolboxes of Pavement Management Systems via Integration of Computational Intelligence*, A Doctoral Dissertation in Civil Engineering, Pennsylvania State University.

LTPP (2013). *Long-Term Pavement Performance (LTPP) Standard Data Release 27.0*. Federal Highway Administration, U.S. Department of Transportation.

MacKay, D.J.C. (1992). "Bayesian Interpolation," *Neural Computation*, Vol. 4, pp. 415-447.

Perera, R. W. and Kohn, S. D. (2001). *LTPP Data Analysis: Factors Affecting Pavement Smoothness.* NCHRP Web Document 40. National Cooperative Highway Research Program, Transportation Research Board.

Von Quintus, H. L., Eltahan, A., and Yau, A. (2001). "Smoothness Models for Hot-Mix Asphalt-Surfaced Pavements; Developed from Long-Term Pavement Performance Data," *Transportation Research Record*, No. 1764, pp. 139-156.