Distribution Comparison for Site-Specific Regression Modeling in Agriculture*

Dragoljub Pokrajac[1], Tim Fiez[2], Dragan Obradovic[3] , Stephen Kwek[1] and Zoran Obradovic[1]
{dpokraja, tfiez, kwek, zoran}@eecs.wsu.edu dragan.obradovic@mchp.siemens.de

[1]School of Electrical Engineering and Computer Science and [2]Department of Crop and Soil Sciences
Washington State University, Pullman, WA 99164, USA
[3]Siemens AG, Corporate Technology, Information and Communications,
Otto-Hahn-Ring 6, 81739 Munich, Germany

Abstract

A novel method for problem decomposition and for local model selection in a multi-model prediction system is proposed. The proposed method partitions the data into disjoint subsets obtained by the local regression modeling and then it learns the distributions on these sets in order to identify the most appropriate regression model for each test point. The system is applied to a site-specific agriculture domain and is shown to provide a substantial improvement in the prediction quality as compared to a global model. Also, some aspects of local learner choice and setting of their parameters are discussed and an overall ability of the proposed model to accurately perform regression is assessed.

Purpose

Technological advances such as the global positioning system and computer controlled variable rate application equipment are enabling agriculture producers to vary rates of fertilizers and other crop production inputs within fields [7]. To derive application maps, producers are collecting large amounts of site specific data such as soil fertility levels and previous crop yields. Most often, these site-specific data are used as input into existing general agronomic recommendation models. These models usually consider only one or a few variables at a time and often have been developed for the "typical" conditions of a fairly large agricultural region (the Eastern Washington for example). In many cases, general non-site specific recommendation models are all that exist. Unfortunately, recent studies [7] show that traditional crop production recommendation models cannot be scaled to a site-specific basis. To adequately predict site-specific crop production needs requires both site-specific data and site-specific recommendation models.

One promising way to develop site-specific recommendations is to learn site-specific models from site-specific data sets containing important driving variables and crop yield [7]. If adequate yield response functions can be defined, optimum production input levels can then be calculated. The purpose of this work is to develop a procedure for defining locally specialized regression models and determining the most appropriate model for a given site-specific data vector.

Method

One of the primary premises of site-specific agriculture is that fields are heterogeneous [7]. Therefore, it follows that multiple, locally specialized models may be better suited for site-specific yield prediction than a single global model. However, with the development of multiple models, one must also develop methods to determine which model to apply for any given pattern not used in the training process. These might be test patterns or any new data for which predictions are desired.

Lazarevic et al. has recently presented one approach to the development and selection of locally specialized models for site-specific agriculture [8]. They merged multiple fields to identify a set of spatial clusters using parameters that should influence crop yield but not the yield itself. Three yield prediction models were then fit to each cluster in a training portion of the merged field

*Partial support by the INEEL University Research Consortium project No.C94-175936 to T. Fiez and Z. Obradovic is gratefully acknowledged.

data. The three models were for low, average, and high yield classes. For each point in the test set, its corresponding cluster is identified. Then, the nearest point from the training set which belongs to the same cluster is found and the corresponding regression model is applied.

Here, a different approach for developing a sequence of local regression models each having a good fit on particular subsets of training data, is considered. This approach is depicted in the following figure:

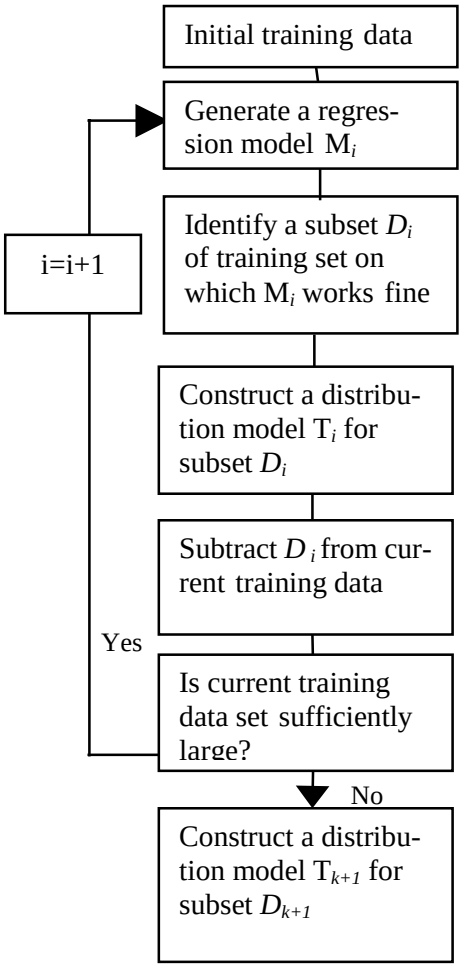

Figure 1: Block-diagram of the proposed method for learning local regression models

After training local models, the corresponding distribution functions are learned. Each training data point is then assigned to the distribution it fits the best and the corresponding local model is used to predict yield.

In the training phase (Fig. 1), we first construct regression models M_i i=0,…k, each result providing superior for a subset of the training data. These can be either local linear regression models [5] or local feedforward neural networks [6] trained on disjoint data sets D_i, $i=1,2,\ldots,k$ The initial data set, denoted by D, consists of all training data provided to the learner and the first regression model, denoted as M_1, is trained on D. Then, a subset D_1 of D where M_1 has sufficiently low error, is selected. Similarly, the successive models M_{i+1} are trained on D\ ($D_1 \cup D_2 \cup \ldots \cup D_i$) and corresponding low error subsets D_{i+1} are identified. Finally, the subset D_{k+1} that contains data on which training errors from $M_1,\ldots M_k$ are all too high, is obtained. Data from D_{k+1} are not suitable for regression by any generated model.

The next step is to learn the distribution for each disjoint subset. To accomplish this, a feedforward neural network consisting of *m* nodes in the first and fourth hidden layer and *n* neurons in the second and third hidden layer (Fig. 2) is applied. Attributes of each pattern are used both as inputs and reference outputs to the network. The backpropagation algorithm [1] is used for training. One network T_i is trained with each subset D_i,$i=1,\ldots k+1$. Once the distribution model is trained (when the error on its outputs is sufficiently small), the network T_i contains the distribution information of dataset D_i.

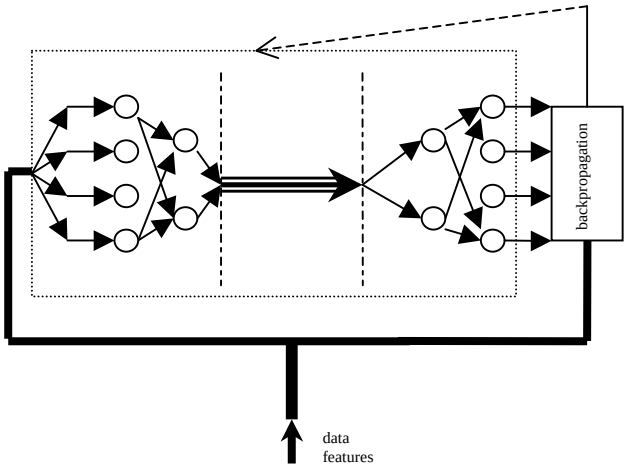

Figure 2: Multiple-layer neural network for data distribution learning

In the testing phase (Fig. 3), for each test pattern x its attributes are assigned to all distribution models to find the network T_i with the smallest error on its output. This implies that pattern x expresses the highest similarity to data set D_j on which T_j is trained. If j=k+1, x is the most similar to dataset D_{k+1} on which no one of trained models M_i, i=1,..k, performs well. Therefore, in such a case, no prediction should be provided. Otherwise, the corresponding model M_i is used to predict on pattern x.

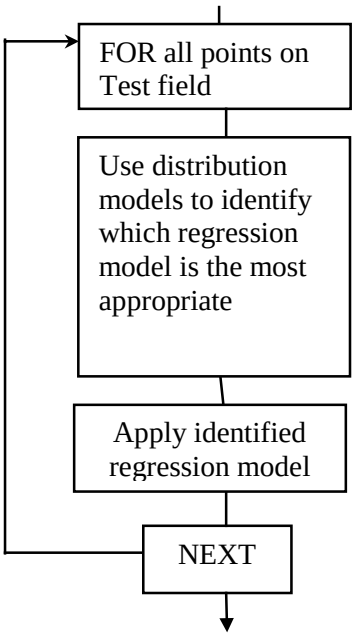

Figure 3: Block-diagram of the application of learned regression models

Observe that even in the case k=1 the herein proposed approach is different from using a simple global model since if there is no evidence that the observed test point x is similar to the part of the training set where model M_1 performs well, the proposed method will not provide a prediction, whereas a global model will always perform regression.

Finally, it is possible that the size of the set D_i of points on which model M_i works fine is small. In this case, model M_i would perform well only on a few points and would add little to overall prediction accuracy. Furthermore, it is difficult to properly learn the distribution D_i given a low number of training patterns. Therefore, a *threshold* for the size of D_i is introduced and only models with $|D_i| > threshold$ are considered.

Results

The proposed methodology was tested on a site-specific agricultural data set obtained from the Idaho National Engineering and Environmental Laboratory SST4Ag Project [7]. The data set contained 7036 patterns on a 10 m x 10 m grid covering a 280 ha spring wheat field. Each pattern contained a x- and y-coordinate and the following soil attributes: salinity, cation exchange capacity (CEC), pH and the concentration of: organic matter, boron, calcium, copper, iron, potassium, magnesium, manganese, nitrate nitrogen, sodium, organic nitrogen, phosphorus, sulfur, and zinc.

The soil attributes were obtained by low-resolution sampling (20 samples from the field). The data from these original sample points were then interpolated to a 10 m x 10 m grid using the inverse distance method [4]. The wheat yield data were obtained using a commercially available combine mounted yield monitor that provided georeferenced yield measurements at 1 to 2 second intervals. These data were interpolated to the same 10 m x 10 m grid as the soil data. To identify a subset of relevant attributes, feature selection using inter-class and probabilistic selection criterion with Mahalanobis distance and branch and bound search was applied to the complete data set [9]. The feature selection procedures indicated that CEC, iron, manganese and salinity were the most useful features and thus these features were used for all regression experiments. To allocate training and testing data, the experimental field was split into two subfields by a north to south dividing line so that each subfield contained 3518 patterns. The East subfield was used for training, and West subfield was used for testing (Fig. 4).

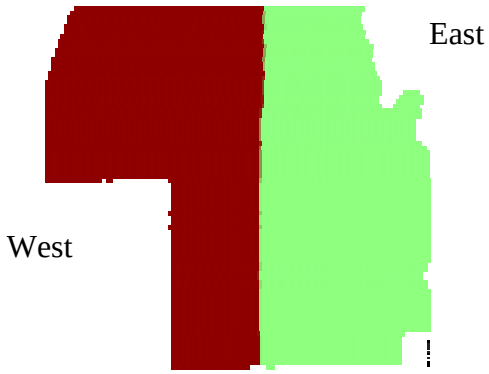

Figure 4: The split of field onto training and test set

Mean wheat yield, the response variable, was similar (about 7% difference) between the training and testing subfields. However, the yield variance of the testing subfield was 54% greater than that of the training subfield, which implies that a single model, once well learned on the training set, can be expected to perform worse on the test.

Neural networks are used in our experiments as regression models M_i. Two-layer perceptrons with 5 non-linear sigmoidal units in a single hidden layer are used and the backpropagation training algorithm is employed [11].
The criterion for partitioning training data into disjoint sets D_i was a squared prediction error less than a threshold. If the prediction error on a pattern was smaller than $\theta=\alpha\sigma^2_{Train}$ where α is a user defined constant, and σ^2_{Train} is the variance on training set, then the pattern was assigned to a set D_i.

To determine an appropriate range for the values of coefficient α, and also to verify the basic assumption that the proposed distribution models can learn the distributions well, we first experimented with one model (k = 1). The model M_1 was trained and points D_1 on which the model performed well and points $D \setminus D_1$ on which it provided unacceptable results were identified. The threshold value α was varied and the mean standard error on the test set was compared with that obtained by applying a global model. Recall that in the case of proposed method, M_1 is applied only on those test points where T_1 has smaller error than T_2, whereas in the case of using global model M_1 is applied on all test points.

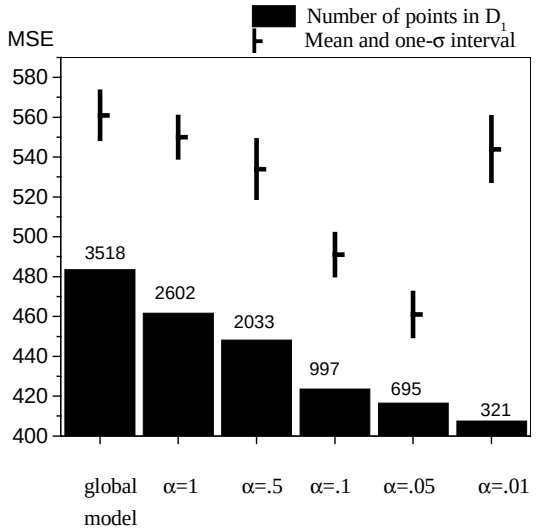

Figure 5: The influence of the error threshold on the size of set D_1 and on the error on the test set if proposed method for k=1 is applied

The results of these experiments, shown in Fig. 5, suggest that with the proper choice of the threshold, network T_1 can identify test data where model M_1 works fine. If a small threshold (α=0.01) is used, the size of set D_1 was insufficient for distribution learning. It resulted in a highly specialized network T_1. In addition, network T_2 trained on heterogeneous data has a tendency to misclassify test data more often. Therefore, some points were wrongly assigned to model M_1 and the consequence was high a mean square error (MSE), sometimes even higher than when only one (global) model was applied. On the other hand, if the threshold is set too high (α=1), subset D_1 was still heterogeneous and network T_1 was not able to train properly and therefore some test points were again misclassified.

The minimum size for sets D_i was also examined. For very small sets D_i, the models, M_i, were too specialized (and appropriate only for very small sets of data), and hence the consequence was the same as if using a small value for α. Selecting a very large minimum size for sets D_i can also result in poor algorithm performance. In this case, the data were too heterogeneous and a specialized regression model, which should perform well on large subset of data, was difficult to develop. Based on informal trials, the minimum size threshold was set to 100 for all experiments

The second phase of experimenting considered an increased number of models. For five models (k=5), an MSE of 316 was obtained which was much better than the MSE of 561 obtained by applying a single conventional model (Table 1).

	global model	k=1	k=5
Number of points on which regression is performed	3518	2602	2033
MSE	561	413	316

Table 1: Effect of using multiple locally specialized models on decreasing MSE

Our experiments show that model parameters must be carefully chosen. For example, we increased the number of hidden neurons in the neural network regression models, M_i, from 5 as used in the previous experiment to 8 and repeated the comparison of one global model versus 5 local models. With 8 hidden neurons, the MSE of the global model was 644, while the MSE when 5 locally specialized models were combined was 544. It appears that increasing the number of hidden neurons is leading to overfitting [6] and resultant poor performance on the test set. Overfitting in our approach can be controlled by the complexity of the local model and by the threshold for the minimum set size of D_i. The probability of overfitting is proportional to the complexity of the local model (number of parameters) and inversely proportional to the minimum size of D_i.

The question arises about the optimal number of locally specialized models. If the number of models M_i is too large, then their corresponding D_is will decrease in size and the models will be increasingly specialized. Hence, the distribution-estimating networks, T_is, will have more difficulties learning the distribution, and the probability that a test point will be assigned to the appropriate model will be smaller. In our experiments, after 5 models were evaluated, subsequent training of models did not result in sufficiently large sets D_i and therefore training was halted.

An important consideration in our method for developing locally specialized regression models is the choice of

model type to employ for M_i. Generally, it is desirable to use regression models with enough expressive power, since with increasing *i*, models M_i should gradually learn more "difficult" distributions. Performed experiments suggest that if M_1 is a statistical linear regression model, subsequent data sets D_i on which M_i are trained, are not large enough. Therefore for M_i, it seems that linear regression models are not suitable. However, the parameters of linear statistical models are deterministically calculated from the training data while the random character of the neural network training process leads to instability of trained models and hence sets D_i. One obvious way to make D_i more stable but to retain expressive power might be to average an ensemble of neural network models (e.g. through bagging [2]) instead of using a single M_i and research towards such a goal is in progress.

While experimentation has shown that our method for developing and selecting locally specialized models can result in better performance than global models, we are currently studying several issues that could lead to further performance improvements.

First, observe that in order for our method to perform well, for each test point $x \in D_i$ distribution model T_i should have the minimal error among all models T_j, j=1,..k. Because models T_i are trained to minimize average error on distributions D_i, there can be subregions $S \subset D_i$ on which T_i performs poorly as compared to the other distribution models. In these subregions, the choice of M_i based on the T_i that gives the smallest error will be incorrect leading to poor predictions. It is possible that employing simultaneous instead of successive training of distribution models could avoid this problem.

Secondly, observe that each model M_i is trained on points for which none of the previous models perform well. However, this does not imply that model M_i would not perform well on subsets D_{i-1}, $D_{i-2...}$ D_1 on which previous models performed well. Therefore, assuming that T_i properly learns D_i, there is an asymmetry of the usage of these networks when determining an optimal model for a test point. Namely, although some regression model M_i chosen for regression on a test point performs well, there is possibility that some other model M_j j>i, performs better. This asymmetry is not considered in the current implementation. One of the ways to alleviate this is to apply weighted combinations of regression models instead of the one most appropriate local model, where weights are inversely proportional to the errors of the distribution models.

New aspect of work

A novel method for problem decomposition and for local model selection in a multi-model prediction system is proposed. It is demonstrated that the proposed approach can provide substantial improvements in the prediction quality as compared to using a single global model on all test data. In addition, parameter selection for the proposed method is discussed and further refinements are suggested.

Conclusions

This work proposed a promising method for increasing yield prediction accuracy in site-specific agriculture. Identification of proper modeling parameters in the proposed system is currently performed by an expensive trial and error process. The aim of our research in progress is to identify a more efficient and computationally stable technique for determining values of these parameters in large scale spatially distributed databases.

References

[1] Bishop, C., *Neural networks for pattern recognition*, Oxford, 1994.

[2] Breiman, L., "Bagging predictors," *Machine learning*, vol. 24, pp 123-140, 1996.

[3] Breiman, L., at al, *Classification and regression trees*, Chapman&Hall, 1984.

[4] Cressie, N., *Statistics for spatial data*, John Wiley and Sons, 1993.

[5] Devore, J.L., *Probability and statistics for engineering and sciences*, Duxbury Press, 1995.

[6] Haykin, S., *Neural networks, a comprehensive foundation*, Prentice Hall, 1998.

[7] Hoskinson, R.L., Hess, J.R., Hempstead, D.W., "Precision farming results from using the decision support system for agriculture (DSS4AG),"*Proc. of the First Int'l Conf. on Geospatial Inf. in Agric. and Forestry, vol. I*, Lake Buena Vista, Florida, June 1-3, 1998, pp. 206-210.

[8] Lazarevic, A., Xu, X., Fiez, T. and Obradovic, Z., "Clustering-regression-ordering steps for knowledge discovery in spatial databases," *Proc. IEEE/INNS Int'l Conf. on Neural Neural Networks*, Washington, D.C., July 1999, in press.

[9] Liu, N., Motoda, H., *Feature selection for knowledge discovery and data mining*, Kluwer Academic Publishing, 1998.

[10] Richards, J., *Remote sensing digital image analysis*, Springer Verlag, 1986.

[11] Werbos, P., *Beyond Regression: New tools for predicting and analysis in the behavioral sciences*, Harvard University, Ph.D. Thesis,1974. Reprinted by Wiley and Sons, 1995.