Using Advanced Regression Models for Determining Optimal Soil Heterogeneity Indicators

Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner

Abstract Nowadays in agriculture, with the advent of GPS-based vehicles and sensor-aided fertilization, large amounts of data are collected. With the importance of carrying out effective and sustainable agriculture getting more and more obvious, those data have to be turned into information – clearly a data analysis task.

Furthermore, there are novel soil sensors which might indicate a field's heterogeneity. Those sensors have to be evaluated and their potential usefulness should be assessed. Our approach consists of two stages, of which the first stage is presented in this article.

The data attributes will be comparable to the ones described in Ruß (2008). In the first stage, we will build and evaluate models for the given data sets. We will present a comparison between results using neural networks, regression trees and SVM regression. Results for an MLP neural network have been published in Ruß et al. (2008). In a future second stage, we will use the model information to evaluate and classify new sensor data. We will then assess their usefulness for the purpose of (yield) optimization.

1 Introduction

Due to the modernization and better affordability of state-of-the-art GPS technology and a multitude of available sensors, a farmer nowadays harvests not only crops but also growing amounts of data. These data are small-scale and precise – which is essentially why the combination of GPS, agriculture and data has been termed *precision agriculture*.

However, collecting large amounts of data often is both a blessing and a curse. There is a lot of data available containing information about a certain asset – here: soil and yield properties – which should be used to the farmer's advantage. This is a

G. Ruß (✉)
Otto-von-Guericke-Universität Magdeburg, Germany
e-mail: russ@iws.cs.uni-magdeburg.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_50, © Springer-Verlag Berlin Heidelberg 2010

common problem for which the term *data mining* or *data analysis* has been coined. Data analysis techniques aim at finding those patterns or information in the data that are both valuable and interesting to the farmer.

A common specific problem that occurs is yield prediction. As early into the growing season as possible, a farmer is interested in knowing how much yield he is about to expect. In the past, this yield prediction has usually relied on farmers' long-term experience for specific fields, crops and climate conditions. What if a computational model could be generated that allows to predict current year's yield based on past data and current year data? This problem of yield prediction encountered here is one which intelligent data analysis should be applied to. More specifically, multi-dimensional regression techniques could be used for yield prediction.

Nowadays, we can collect small-scale, precise data in-season using a multitude of sensors. These sensors essentially aim to measure a field's heterogeneity. In future work, these sensors will be assessed as to how useful they are for the purpose of yield prediction. For this work, this article should serve as an overview on the capabilities of different regression techniques used on agricultural yield data.

1.1 Research Target

The overall research target is to find those indicators of a field's heterogeneity which are suited best to be used for a yield prediction task. Since this should be done in-season, the sub-task here is one of multi-dimensional regression – predicting yield from past and in-season attributes. At a later stage, when multi-year data are available, models from past years could be used to predict present year's yield.

Therefore, this work aims at finding suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities. Since multi-year data are not yet available, the prediction can only be done in space using cross-validation, instead of in time. As soon as multi-year data are available, the models can be trained using these data for prediction in time. We will evaluate different types of regression techniques on different data sets. Since these models usually are strongly parameterized, an additional question is whether the model parameters can be carried over from one field to other fields which are comparable in (data set) size. This issue will also be addressed in this work. This is especially useful when new data have to be evaluated using one of the presented models.

1.2 Article Structure

Section 2 lays out the data sets that this work builds upon. The attributes and their properties will be presented shortly. Section 3 briefly presents four selected regression techniques from the data mining area which will be used for yield prediction.

Section 4 shows the results of the modeling/regression stage and provides answers to the aforementioned research questions.

At the end of this article, future work is pointed out and implementation details are provided.

2 Data Description

The data available in this work have been obtained in the years 2003–2006 on three fields near Köthen, north of Halle, Germany (GPS coordinates: Latitude N 51 40.430, Longitude E 11 58.110). All information available for these 65-, 72- and 32-hectare fields was interpolated using kriging (Stein 1999) to a grid with 10 by 10 meters grid cell sizes. Each grid cell represents a record with all available information. During the growing season of 2006, the latter field was subdivided into different strips, where various fertilization strategies were carried out. For an example of various managing strategies, see e.g. Schneider and Wagner (2006), which also shows the economic potential of PA technologies quite clearly. The fields grew winter wheat, where nitrogen fertilizer was distributed over three application times during the growing season.

Overall, for each field there are seven attributes – accompanied by the respective current year's yield (2004 or 2006) as the target attribute. Those attributes have been described in detail in Ruß et al. (2008), an overview is given below. In total, for the F04 field there are 5241 records, for F131 there are 2278 records, for F330 there are 4578 records, thereof none with missing values and none with outliers. In addition, a subset for F131 was available: in this subset, a special fertilization strategy was carried out which used a neural network for prediction and optimization – this data set is called F131net and has 1144 records.

In this work, data sets from three different fields are evaluated. A brief summary of two of the available data sets is given in Tables 1a and 1b. On each field, different fertilization strategies have been used. One of those strategies is based on a technique that uses a multi-layer perceptron (MLP) for prediction and optimization. This technique has been presented and evaluated in, e.g., Ruß et al. (2008); Ruß (2008) or Weigert (2006). For each field, one data set will contain all records, thus containing all the different fertilization strategies. In addition, a subset of F131 has been chosen to serve as a fourth data set to be evaluated.

3 Advanced Regression Techniques

As mentioned in the introduction, the task of yield prediction is essentially a task of multi-dimensional regression. Therefore, this section will serve as an overview about different regression techniques that are applicable to the yield data sets. We aim to evaluate these techniques on the data sets presented in the preceding section.

Table 1 Overview of the *F04* and *F131* data sets. The additional data sets *F330* and *F131net*, which is a subset of *F131*, are not shown as their statistical properties are very similar to those of *F04* and *F131*

(a) Data overview, F04					(b) Data overview, F131				
F04	*min*	*max*	*mean*	*std*	**F131**	*min*	*max*	*mean*	*std*
YIELD03	1.19	12.38	6.27	1.48	YIELD05	1.69	10.68	5.69	0.93
EM38	17.97	86.45	33.82	5.27	EM38	51.58	84.08	62.21	8.60
N1	0	100	57.7	13.5	N1	47.70	70	64.32	6.02
N2	0	100	39.9	16.4	N2	14.80	100	51.71	15.67
N3	0	100	38.5	15.3	N3	0	70	39.65	13.73
REIP32	721.1	727.2	725.7	0.64	REIP32	719.6	724.4	722.6	0.69
REIP49	722.4	729.6	728.1	0.65	REIP49	722.3	727.9	725.8	0.95
YIELD04	6.42	11.37	9.14	0.73	YIELD06	1.54	8.83	5.21	0.88

The regression task can be formalized as follows: the training set

$$T = \{\{x_1, \ldots, x_n\}, y_i\}_{i=1}^{N} \tag{1}$$

is considered for the training process, where $x_i, i = 1, \ldots, n$ are continuous input values and $y_i, i = 1 \ldots, N$ are continuous output values. Given this training set, the task of the regression techniques is to approximate the underlying function sufficiently well.

3.1 Introduction to Regression Techniques

Since one particular technique, namely MLPs, has been used successfully in previous work (Ruß et al. 2008; Ruß 2008), it is used as a reference model here. Three additional modeling techniques, namely RBF networks, regression trees, and support vector regression, will be presented, which are suitable for the task of yield prediction. The aforementioned techniques have, to the authors' knowledge, not been compared to each other when used with different data sets in the agriculture context. This section presents some of the background for each of the techniques before they will be evaluated in Sect. 4.

3.2 Neural Networks

In previous work multi-layer perceptrons (MLPs), a type of neural networks, have been used for a modeling task (Ruß et al. 2008; Ruß 2008) similar to the one encountered here. Furthermore, neural networks have shown to be quite effective in modeling yield of different crops (Drummond et al. 1998; Serele et al. 2000). The MLP model was established as a reference model against which further regression techniques would have to compete. For a more detailed and formal description of MLP neural networks, it is referred to Hagan (1995) or Haykin (1998). The network layout and the parameters will be given in Sect. 4. In this work, the matlab implementation for the MLP network was used: `newff`.

Furthermore, a different type of neural network, a radial basis function (RBF) network, will be evaluated, since it is well-suited to the regression task. For this network, matlab's `newrb` function has been utilized.

3.3 Regression Tree

Regression as well as decision trees are usually constructed in a top-down, greedy search approach through the space of possible trees (Mitchell 1997). The basic algorithms for constructing such trees are CART (Breiman et al. 1984), ID3 (Quinlan 1986) and its successor C4.5 (Quinlan 1993). The idea here is to ask the question "which attribute should be tested at the top of the tree?" To answer this question, each attribute is evaluated to determine how well it is suited to split the data. The best attribute is selected and used as the test node. This procedure is repeated for the subtrees. For further information on the construction details and possible problems (such as overlearning) the reader is referred to Mitchell (1997). For this work the standard matlab implementation of `classregtree` has been utilized.

3.4 Support Vector Regression

Support Vector Machines (SVMs) are a supervised learning method discovered by Boser et al. (1992). However, the task here is regression, so the focus is on support vector regression (SVR). A more in-depth discussion can be found in Gunn (1998). Given the training set, the goal of SVR is to approximate a linear function $f(x) = \langle w, x \rangle + b$ with $w \in \mathbb{R}^N$ and $b \in \mathbb{R}$. This function minimizes an empirical risk function defined as

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} L_{\varepsilon}(\hat{y} - f(x)), \tag{2}$$

where $L_{\varepsilon}(\hat{y} - f(x)) = \max((|\xi| - \varepsilon), 0)$. $|\xi|$ is the so-called slack variable, which has mainly been introduced to deal with otherwise infeasible constraints of the optimization problem, as has been mentioned in Smola and Schölkopf (1998). By using this variable, errors are basically ignored as long as they are smaller than a properly selected ε. L_{ε} is called ε-insensitive loss function. Other kinds of functions can be used, some of which are presented in Chap. 5 of Gunn (1998). To estimate $f(x)$, a quadratic problem must be solved. See Mejía-Guevara and Kuri-Morales

(2007) for the dual form of this problem. In this work, the SVMtorch implementation from Collobert et al. (2001) has been utilized. Its documentation also points out further details of the SVR process.

3.5 Linear Regression and Naive Estimator

For comparison reasons, two further prediction methods are employed to compare the advanced regression techniques against. The first of these is a simple multilinear regression estimator. The second is a naive estimator which simply reports the previous year's yield as the output yield of the current year.

3.6 Model Parameter Estimation

Each of the aforementioned four different models will be evaluated on the same data sets. One of the research goals here is to establish whether a model which has been used on one data set can be used on a different data set without changing its parameters. This would lead us to believe that comparable fields could use the same prediction model. Hence, the *F04* data set is used to determine the model parameters experimentally. Afterwards, the models are re-trained on the remaining data sets using the settings determined for *F04*. The parameter settings are given in Sect. 4.

For training the models, a cross-validation approach is taken. As mentioned in e.g. Hecht-Nielsen (1990), the data will be split randomly into a training set, a validation set and a test set. The model is trained using the training data and after each training iteration, the error on the validation data is computed. During training, this error usually declines towards a minimum. Beyond this minimum, the error rises – overlearning (or overfitting) occurs: the model fits the training data perfectly but does not generalize well. Hence, the model training is stopped when the error on the validation set starts rising. A size ratio of 8:1:1 for training, validation and test sets is used. The data sets are partitioned randomly 20 times and the models are trained. The models' performance will be determined using the root mean squared error (RMSE) and the mean absolute error (MAE) on the test set. It is assumed that the reader is familiar with these measures.

4 Regression Results

The models are run with the parameter settings given below. Those were determined experimentally on *F04* using a grid search, and carried over to the remaining data sets.

MLP A relatively small number of 10 hidden neurons is used and the network is trained until a minimum gradient of 0.001 is reached, using a learning rate of 0.25 and the *tangens hyperbolicus* sigmoid activation function.

Table 2 Results of running different models on different data sets. The best predictive model for each data set is marked in **bold** font

Model/Dataset	MAE				RMSE			
	F04	F131	F131net	F330	F04	F131	F131net	F330
MLP	0.3706	0.2468	0.2300	0.3576	0.4784	0.3278	0.3073	0.5020
RBF	0.3838	0.2466	0.2404	0.3356	0.5031	0.3318	0.3205	**0.4657**
REGTREE	0.4380	0.2823	0.2530	0.4151	0.5724	0.3886	0.3530	0.6014
SVR	**0.3446**	**0.2237**	**0.2082**	**0.3260**	**0.4508**	**0.3009**	**0.2743**	0.4746
LINREG	0.4285	0.3257	0.2766	0.3820	0.5578	0.4392	0.3871	0.5330
NAIVE	2.9061	0.6135	0.6492	4.7157	3.1253	0.7613	0.7847	4.8308

RBF For the radial basis function network, a radius of 1 is used for the radial basis neurons in the hidden layer. The algorithm, which incrementally adds neurons until the error goal of 0.001 is met, uses a maximum number of 70 neurons.

RegTree For the regression tree, the default settings of classregtree perform optimal; the full tree is pruned automatically and the minimum number of training examples below which no split should be done is 10.

SVR For the support vector regression model, the radial basis function kernel yields the best results, using the parameters $C = 60$, $\sigma = 4.0$ and $\xi = 0.2$.

Considering the results in Table 2, support vector regression obviously performs best on all but one of the data sets, regarding both error measures. Furthermore, SVR also is the model taking the least amount of computation time. Hence, the slight difference between the RMSE of SVR and RBF on the *F330* data set may be considered insignificant in practice when computational cost is also taken into account when deciding for a model. Regarding the understandability of the generated models, it would certainly be desirable to have the regression tree as the best model since simple decision rules can easily be generated from the tree. However, the regression tree performs worst in all of the cases. On the other hand, when comparing the hitherto reference model MLP with the current best model SVR, there is not much difference in the understandability of both models.

5 Conclusion

The results clearly show that support vector regression can serve as a better reference model for yield prediction than MLP. Even if the improvement should be statistically insignificant, the advantages of SVR over MLP remain. It is computationally less demanding, at least as understandable as the MLP and, most importantly, mostly produces better yield predictions. Furthermore, the comparison against a linear regression baseline and a naive estimator shows that the additional effort for using SVR is worth it.

Furthermore, the results also show that model parameters which have been established on one data set can be carried over to different (but similar with respect to the attributes) data sets. A model for identifying the most useful heterogeneity indicators is currently being evaluated.

5.1 Future Work

Due to the relatively high spatial resolution of the input data, the possible issue of spatial autocorrelation arises. This influences the modeling during the cross-validation stage. This will be investigated in future work.

Acknowledgements Experiments have been conducted using Matlab 2008a. The field trial data came from the experimental farm Görzig of Martin-Luther-University Halle-Wittenberg, Germany.

References

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). New York: ACM Press.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Collobert, R., Bengio, S., & Williamson, C. (2001). Svmtorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research, 1*, 143–160.

Drummond, S., Joshi, A., & Sudduth, K. A. (1998). Application of neural networks: precision farming. In *International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence* (Vol. 1, pp. 211–215).

Gunn, S. R. (1998). *Support vector machines for classification and regression.* Technical Report, School of Electronics and Computer Science, University of Southampton, Southampton, U.K.

Hagan, M. T. (1995). *Neural network design (electrical engineering)*. Thomson Learning.

Haykin, S. (1998). *Neural networks: A Comprehensive Foundation* (2nd ed.). Englewood, Cliffs, NJ: Prentice Hall.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Reading, MA, USA: Addison-Wesley.

Mejía-Guevara, I., & Kuri-Morales, A. (2007). Evolutionary feature and parameter selection in support vector regression. In *Lecture Notes in Computer Science* (Vol. 4827, pp. 399–408). Berlin, Heidelberg: Springer.

Mitchell, T. M. (1997). *Machine learning*. NY, USA: McGraw-Hill Science/Engineering/Math.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Los Altos, CA: Morgan Kaufmann.

Ruß, G., Kruse, R., Schneider, M., & Wagner, P. (2008). Estimation of neural network parameters for wheat yield prediction. In M. Bramer (Ed.), *Artificial Intelligence in Theory and Practice II* of *IFIP International Federation for Information Processing* (Vol. 276, pp. 109–118). Berlin: Springer.

Ruß, G., Kruse, R., Schneider, M., & Wagner, P. (2008). Optimizing wheat yield prediction using different topologies of neural networks. In J. L. Verdegay, M. Ojeda-Aciego, & Magdalena, L. (Eds.), *Proceedings of IPMU-08* (pp. 576–582). University of Málaga.

Ruß, G., Kruse, R., Wagner, P., & Schneider, M. (2008). Data mining with neural networks for wheat yield prediction. In P. Perner (Ed.), *Advances in Data Mining (Proc. ICDM 2008)* (pp. 47–56). Berlin, Heidelberg: Springer Verlag.

Schneider, M., & Wagner, P. (2006). Prerequisites for the adoption of new technologies – the example of precision agriculture. In *Agricultural Engineering for a Better World*, Düsseldorf: VDI Verlag GmbH.

Serele, C. Z., Gwyn, Q. H. J., Boisvert, J. B., Pattey, E., Mclaughlin, N., & Daoust, G. (2000). Corn yield prediction with artificial neural network trained using airborne remote sensing and topographic data. In *2000 IEEE International Geoscience and Remote Sensing Symposium, 1*, 384–386.

Smola, A. J., & Schölkopf, B. (1998). *A tutorial on support vector regression.* Technical report, Statistics and Computing.

Stein, M. L. (1999). *Interpolation of Spatial Data : Some Theory for Kriging (Springer Series in Statistics)*. Berlin: Springer.

Weigert, G. (2006). *Data Mining und Wissensentdeckung im Precision Farming - Entwicklung von ökonomisch optimierten Entscheidungsregeln zur kleinräumigen Stickstoff-Ausbringung*. PhD thesis, TU München.