Data Mining with Neural Networks for Wheat Yield Prediction

Georg Ruß[1], Rudolf Kruse[1], Martin Schneider[2], and Peter Wagner[2]

[1] Otto-von-Guericke-University of Magdeburg
[2] Martin-Luther-University of Halle

Abstract. Precision agriculture (PA) and information technology (IT) are closely interwoven. The former usually refers to the application of nowadays' technology to agriculture. Due to the use of sensors and GPS technology, in today's agriculture many data are collected. Making use of those data via IT often leads to dramatic improvements in efficiency. For this purpose, the challenge is to change these raw data into useful information. In this paper we deal with neural networks and their usage in mining these data. Our particular focus is whether neural networks can be used for predicting wheat yield from cheaply-available in-season data. Once this prediction is possible, the industrial application is quite straightforward: use data mining with neural networks for, e.g., optimizing fertilizer usage, in economic or environmental terms.

Keywords: Precision Agriculture, Data Mining, Neural Networks, Prediction.

1 Introduction

Due to the rapidly advancing technology in the last few decades, more and more of our everyday life has been changed by information technology. Information access, once cumbersome and slow, has been turned into "information at your fingertips" at high speed. Technological breakthroughs have been made in industry and services as well as in agriculture. Mostly due to the increased use of modern GPS technology and advancing sensor technology in agriculture, the term *precision agriculture* has been coined. It can be seen as a major step from uniform, large-scale cultivation of soil towards small-field, precise planning of, e.g., fertilizer or pesticide usage. With the ever-increasing amount of sensors and information about their soil, farmers are not only harvesting, e.g., potatoes or grain, but also harvesting large amounts of data. These data should be used for optimization, i.e. to increase efficiency or the field's yield, in economic or environmental terms.

Until recently [13], farmers have mostly relied on their long-term experience on the particular acres. With the mentioned technology advances, cheaper sensors have eased data acquisition on such a scale that it makes them interesting for the data mining community. For carrying out an information-based field cultivation,

P. Perner (Ed.): ICDM 2008, LNAI 5077, pp. 47–56, 2008.

the data have to be transformed into utilizable information in terms of management recommendations as a first step. This can be done by decision rules, which incorporate the knowledge about the coherence between sensor data and yield potential. In addition, these rules should give (economically) optimized recommendations. Since the data consist of simple and often even complete records of sensor measurements, there are numerous approaches known from data mining that can be used to deal with these data. One of those approaches are artificial neural networks [4] that may be used to build a model of the available data and help to extract the existing pattern. They have been used before in this context, e.g. in [1], [7] or [12].

The connection between information technology and agriculture is and will become an even more interesting area of research in the near future. In this context, IT mostly covers the following three aspects: data collection, analysis and recommendation [6]. This work is based on a dissertation that deals with data mining and knowledge discovery in precision agriculture from an agrarian point of view [15]. This research led to economically optimized decision rules, but left out some of the details on the used techniques. Since we are dealing with the above-mentioned data records, the computer science perspective will be applied. The main research target is whether we can model and optimize the site-specific data by means of further computational intelligence techniques. We will therefore deal with data collection and analysis.

The paper is structured as follows: Section 2 will provide the reader with details on the acquisition of the data and some of the data's properties. Section 3 will give some background information on neural networks. In Section 4 we will describe the experimental layout and afterwards, we will evaluate the results that were obtained. The last section will give a brief conclusion.

2 Data Acquisition

The data available in this work have been obtained in the years 2003 and 2004 on a field near Köthen, north of Halle, Germany. All information available for this 65-hectare field was interpolated 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 2004, the field was subdivided into different strips, where various fertilization strategies were carried out. For an example of various managing strategies, see e.g. [11], which also shows the economic potential of PA technologies quite clearly. The field grew winter wheat, where nitrogen fertilizer was distributed over three application times.

Overall, there are seven input attributes – accompanied by the yield in 2004 as the target attribute. Those attributes will be described in the following. In total, there are 5241 records, thereof none with missing values and none with outliers.

2.1 Nitrogen Fertilizer – N1, N2, N3

The amount of fertilizer applied to each subfield can be easily measured. It is applied at three points in time into the vegetation period. Since the site of

Table 1. Data overview

Attribute	*min*	*max*	*mean*	*std*	*Description*
N1	0	100	57.7	13.5	amount of nitrogen fertilizer applied at the first date
N2	0	100	39.9	16.4	amount of nitrogen fertilizer applied at the second date
N3	0	100	38.5	15.3	amount of nitrogen fertilizer applied at the third date
REIP32	721.1	727.2	725.7	0.64	red edge inflection point vegetation index
REIP49	722.4	729.6	728.1	0.65	red edge inflection point vegetation index
EM38	17.97	86.45	33.82	5.27	electrical conductivity of soil
Yield03	1.19	12.38	6.27	1.48	yield in 2003
Yield04	6.42	11.37	9.14	0.73	yield in 2004

application had also been designed as an experiment for data collection, the range of N1, N2, and N3 in the data is from 0 to 100 $\frac{kg}{ha}$, where it is normally at around $60\frac{kg}{ha}$.

2.2 Vegetation – REIP32, REIP49

The *red edge inflection point* (REIP) is a first derivative value calculated along the red edge region of the spectrum, which is situated from 680 to 750nm. Dedicated REIP sensors are used in-season to measure the plants' reflection in this spectral band. Since the plants' chlorophyll content is assumed to highly correlate with the nitrogen availability (see, e.g. [10]), the REIP value allows for deducing the plants' state of nutrition and thus, the previous crop growth. For further information on certain types of sensors and a more detailed introduction, see [15] or [8]. Plants that have less chlorophyll will show a lower REIP value as the red edge moves toward the blue part of the spectrum. On the other hand, plants with more chlorophyll will have higher REIP values as the red edge moves toward the higher wavelengths. For the range of REIP values encountered in the available data, see Table 1. The numbers in the REIP32 and REIP49 names refer to the growing stage of winter wheat.

2.3 Electric Conductivity – EM38

A non-invasive method to discover and map a field's heterogeneity is to measure the soil's conductivity. Commercial sensors such as the EM-38[1] are designed for agricultural use and can measure small-scale conductivity to a depth of about 1.5 metres. There is no possibility of interpreting these sensor data directly in terms of its meaningfulness as yield-influencing factor. But in connection with other site-specific data, as explained in the rest of this section, there could be coherences. For the range of EM values encountered in the available data, see Table 1.

[1] Trademark of Geonics Ltd, Ontario, Canada.

Table 2. Overview on available data sets for the three fertilization times (FT)

FT1	Yield03, EM38, N1
FT2	Yield03, EM38, N1, REIP32, N2
FT3	Yield03, EM38, N1, REIP32, N2, REIP49, N3

2.4 Yield 2003/2004

Here, yield is measured in $\frac{t}{ha}$. In 2003, the range for corn was from 1.19 to 12.38. In 2004, the range for wheat was from 6.42 to 11.37, with a higher mean and smaller standard deviation, see Table 1.

2.5 Data Overview

A brief summary of the available data attributes is given in Table 1.

2.6 Points of Interest

From the agricultural perspective, it is interesting to see how much the influencable factor "fertilization" really determines the yield in the current site-year. Furthermore, there may be additional factors that correlate directly or indirectly with yield and which can not be discovered using regression or correlation analysis techniques like PCA. To determine those factors we could establish a model of the data and try to isolate the impact of single factors. That is, once the current year's yield data can be predicted sufficiently well, we can evaluate single factors' impact on the yield.

From the data mining perspective, there are three points in time of fertilization, each with different available data on the field. What is to be expected is that, as more data is available, after each fertilization step the prediction of the current year's yield (`Yield03`) should be more precise. Since the data have been described in-depth in the preceding sections, Table 2 serves as a short overview on the three different data sets for the specific fertilization times.

For each data set, the `Yield04` attribute is the target variable that is to be predicted. Once the prediction works sufficiently well and is reliable, the generation of, e.g., fertilization guidelines can be tackled. Therefore, the following section deals with an appropriate technique to model the data and ensure prediction quality.

3 Data Modeling

In the past, numerous techniques from the computational intelligence world have been tried on data from agriculture. Among those, neural networks have been quite effective in modeling yield of different crops ([12], [1]). In [14] and [15], artificial neural networks (ANNs) have been trained to predict wheat yield from fertilizer and additional sensor input. However, from a computer scientist's perspective, the presented work omits details about the ANN's internal settings,

such as network topology and learning rates. In the following, an experimental layout will be given that aims to determine the optimal parameters for the ANN.

3.1 Neural Networks Basics

The network type which will be optimized here are multi-layer perceptrons (MLPs) with backpropagation learning. They are generally seen as a practical vehicle for performing a non-linear input-output mapping [4]. To counter the issue of overfitting, which leads to perfect performance on training data but poor performance on test or real data, cross-validation will be applied. As mentioned in e.g. [5], the data will be split randomly into a training set, a validation set and a test set. Essentially, the network will be trained on the training set with the specified parameters. Due to the backpropagation algorithm's properties, the error on the training set declines steadily during the training process. However, to maximize generalization capabilities of the network, the training should be stopped once the error on the validation set rises [2].

As explained in e.g. [3], advanced techniques like Bayesian regularization [9] may be used to optimize the network further. However, even with those advanced optimization techniques, it may be necessary to train the network starting from different initial conditions to ensure robust network performance. For a more detailed and formal description of neural networks, we refer to [3] or [4].

3.2 Variable Parameters

For each network there is a large variety of parameters that can be set. However, one of the most important parameters is the network topology. For the data set described in Section 2, the MLP structure should certainly have up to seven input neurons and one output neuron for the predicted wheat yield. Since we are dealing with more than 5000 records, the network will require a certain amount of network connections to be able to learn the input-output mapping sufficiently well. Furthermore, it is generally unclear and mostly determined experimentally how many layers and how many neurons in each layer should be used [2]. Therefore, this experiment will try to determine those network parameters empirically. Henceforth, it is assumed that two layers are sufficient to approximate the data set. This structure is generally assumed to be capable of approximating virtually any function of interest, provided that sufficiently many hidden connections are available [5]. To determine the exact number of neurons, a maximum size of 32 neurons in the first and second hidden layer has been chosen – this provides a maximum of 1024 connections in between the hidden layers, which should be sufficient. The range of the network layers' sizes will be varied systematically from 2 to 32. The lower bound of two neurons has been chosen since one neuron with a sigmoidal transfer function does not contribute much to the function approximation capabilities. The upper bound is generally problem-dependent; here, it was determined by preliminary experiments that showed that the generalization capabilities are reduced by using more than a certain number of neurons. Moreover, the maximum network size has also been chosen for reasons of computation time.

3.3 Fixed Parameters

In preliminary experiments which varied further network parameters systematically, a learning rate of 0.5 and a minimum gradient of 0.001 have been found to deliver good approximation results without overfitting the data. All of the network's neurons have been set to use the *tanh* transfer function, the initial network weights have been chosen randomly from an interval of $[-1, 1]$. Data have been normalized to an interval of $[0, 1]$.

3.4 Network Performance

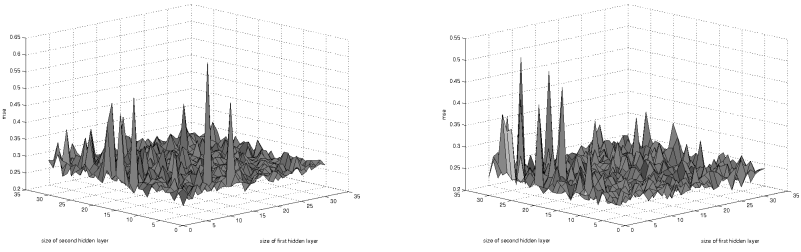
The network performance with the different parameters will be determined by the mean of the squared errors on the test set since those test data will not be used for training. Overall, there are three data sets for which a network will be trained. The network topology is varied from 2 to 32 neurons per layer, leaving 961 networks to be trained and evaluated. The network's approximation quality can then be shown on a surface plot.

4 Results and Discussion

To visualize the network performance appropriately, a surface plot has been chosen. In each of the following figures, the x- and y-axes show the sizes of the first and second hidden layer, respectively. Figures 1(a), 1(b) and 2(a) show the mean squared error vs. the different network sizes, for the three fertilization times (FT), respectively. For the first FT, the mse on average is around 0.3, at the second FT around 0.25 and at the third FT around 0.2. It had been expected that the networks' prediction improves once more data (in terms of attributes) become available for training. There is, however, no clear tendency towards better prediction with larger network sizes. Nevertheless, a prediction accuracy of between 0.44 and 0.55 $\frac{t}{ha}$ (the figures only show the mean *squared* error) at an average yield of 9.14 $\frac{t}{ha}$ is a good basis for further developments with those data and the trained networks.

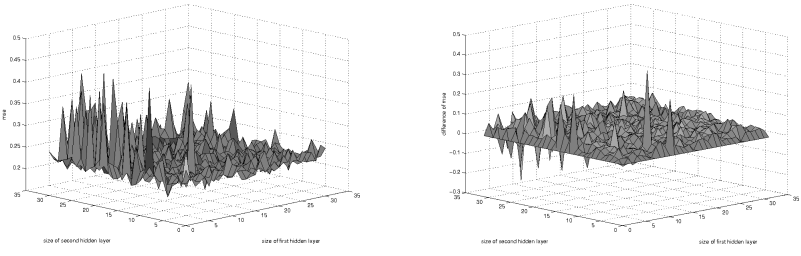
Furthermore, there are numerous networks with bad prediction capabilities in the region where the first hidden layer has much fewer neurons than the second hidden layer. Since we are using feedforward-backpropagation networks without feedback, this behaviour should also be as expected: the information that leaves the input layer is highly condensed in the first hidden layer if it has from two to five neurons – therefore, information is lost. The second hidden layer's size is then unable to contribute much to the network's generalization – the network error rises.

For the choice of network topology, there is no general answer to be given using any of the data sets from the different FTs. What can be seen is that the error surface is quite flat so that a layout with 16 neurons in both hidden layers should be an acceptable tradeoff between mean squared error and computational complexity.



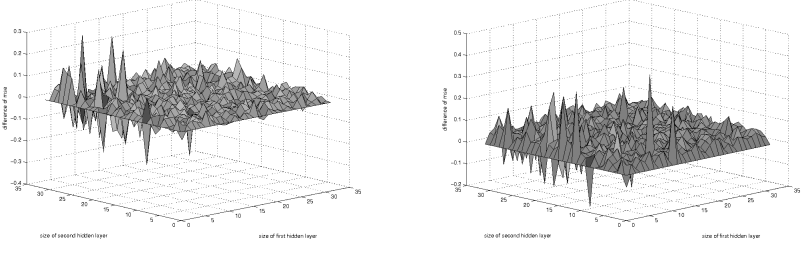
(a) MSE for first data set (b) MSE for second data set

Fig. 1. MSE plots for first and second data set



(a) MSE for third data set (b) MSE difference from first to second data set

Fig. 2. MSE plot for third data set, MSE difference plot for first data set



(a) MSE difference from second to third data set (b) MSE difference from first to third data set

Fig. 3. MSE difference plots for second and third data set

4.1 Difference Plots

Figures 2(b), 3(a) and 3(b) show the difference between the networks' mean squared errors vs. the different network sizes, respectively. Therefore, they illustrate the networks' performance quite clearly. In the majority of cases, the networks generated from later data sets, i.e. those with more information, can predict the target variable better than the networks from the earlier data sets.

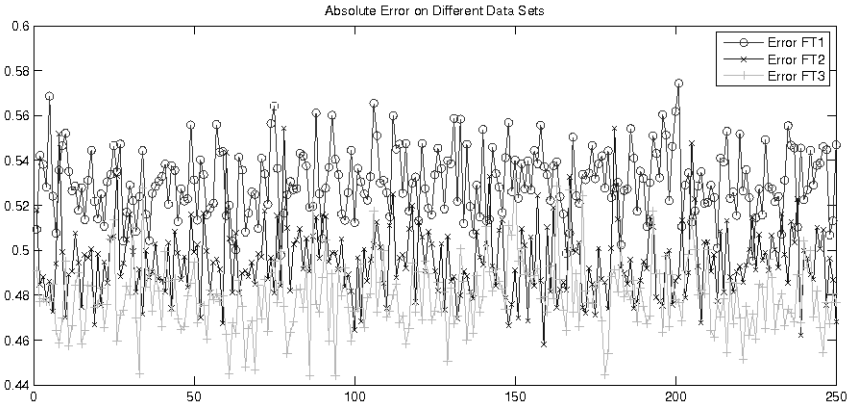


Fig. 4. Comparison of data sets, absolute error vs. trial index

4.2 Comparison of Data Sets FT1, FT2, FT3

In the preceding section we assumed that the networks trained on those data that were available later into the season perform better than the ones on less, earlier data. To substantiate this claim we fixed the network structure to the one that we established earlier: two hidden layers with 16 neurons each and fully connected. The three data sets FT1, FT2, and FT3 were divided randomly into training, validation and testing set at a ratio of 0.6/0.2/0.2. The division and training steps were repeated 250 times and the absolute error was recorded. Figure 4 shows the error on the different data sets against the trial index. It can be seen quite clearly that our assumption could be substantiated: the average error on FT3 is considerably smaller than the one on FT1 or FT2. For FT1, the mean error is 0.53; for FT2, it is 0.49; and for FT3 it is 0.48. The error's standard deviation on all data sets is 0.015.

5 Conclusion

This paper contributes to finding and evaluating models of agricultural yield data. Starting from a detailed data description, we built three data sets that could be used for training. In earlier work, neural networks had been used to model the data. Certain parameters of the ANNs have been evaluated, most important of which is the network topology itself. We built and evaluated different networks and substantiated the assumption that the prediction accuracy of the networks rises once more data become available at later stages into the growing season.

5.1 Future Work

In subsequent work, we will compare ANNs with suitable further techniques (such as regression or SVMs) to find the best predictor. We will make use of

those techniques to model site-year data from different years. It will be evaluated whether the data from one year are sufficient to predict subsequent years' yields. It will also be interesting to study to which extent one field's results can be carried over to modeling a different field. The impact of different parameters during cropping and fertilization on the yield will be evaluated. Finally, controllable parameters such as fertilizer input can be optimized, environmentally or economically.

Acknowledgements

Experiments have been conducted using Matlab 2007b and the corresponding Neural Network Toolbox 5.1. The field trial data came from the experimental farm Görzig of Martin-Luther-University Halle-Wittenberg, Germany. The matlab script that produced Figure 4 and can easily be tailored towards producing the remaining figures can be downloaded from `http://tinyurl.com/2fmk2m` or can otherwise be requested from the first author of this work.

References

1. Drummond, S., Joshi, A., Sudduth, K.A.: Application of neural networks: precision farming. In: The 1998 IEEE International Joint Conference on Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence, vol. 1, pp. 211–215 (1998)
2. Fausett, L.V.: Fundamentals of Neural Networks. Prentice Hall, Englewood Cliffs (1994)
3. Hagan, M.T.: Neural Network Design (Electrical Engineering). Thomson Learning (December 1995)
4. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice Hall, Englewood Cliffs (1998)
5. Hecht-Nielsen, R.: Neurocomputing. Addison-Wesley, Reading (1990)
6. Heimlich, R.: Precision agriculture: information technology for improved resource use. Agricultural Outlook, 19–23 (April 1998)
7. Kitchen, N.R., Drummond, S.T., Lund, E.D., Sudduth, K.A., Buchleiter, G.W.: Soil Electrical Conductivity and Topography Related to Yield for Three Contrasting Soil-Crop Systems. Agron J. 95(3), 483–495 (2003)
8. Liu, J., Miller, J.R., Haboudane, D., Pattey, E.: Exploring the relationship between red edge parameters and crop variables for precision agriculture. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS 2004, vol. 2, pp. 1276–1279 (2004)
9. MacKay, D.J.C.: Bayesian interpolation. Neural Computation 4(3), 415–447 (1992)
10. Middleton, E.M., Campbell, P.K.E., Mcmurtrey, J.E., Corp, L.A., Butcher, L.M., Chappelle, E.W.: "Red edge" optical properties of corn leaves from different nitrogen regimes. In: IEEE International Geoscience and Remote Sensing Symposium, 2002. IGARSS 2002, vol. 4, pp. 2208–2210 (2002)
11. Schneider, M., Wagner, P.: Prerequisites for the adoption of new technologies - the example of precision agriculture. In: Agricultural Engineering for a Better World, Düsseldorf. VDI Verlag GmbH (2006)

12. Serele, C.Z., Gwyn, Q.H.J., Boisvert, J.B., Pattey, E., Mclaughlin, N., Daoust, G.: Corn yield prediction with artificial neural network trained using airborne remote sensing and topographic data. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2000. IGARSS 2000, vol. 1, pp. 384–386 (2000)
13. Sonka, S.T., Bauer, M.E., Cherry, E.T., John, Heimlich, R.E.: Precision Agriculture in the 21st Century: Geospatial and Information Technologies in Crop Management. National Academy Press, Washington (1997)
14. Wagner, P., Schneider, M.: Economic benefits of neural network-generated site-specific decision rules for nitrogen fertilization. In: Stafford, J.V. (ed.) Proceedings of the 6th European Conference on Precision Agriculture, pp. 775–782 (2007)
15. Weigert, G.: Data Mining und Wissensentdeckung im Precision Farming - Entwicklung von ökonomisch optimierten Entscheidungsregeln zur kleinräumigen Stickstoff-Ausbringung. PhD thesis, TU München (2006)