UNDERSTANDING AND VISUALIZING THE DATA MINING PROCESS FOR REAL AGRICULTURE

Dr.V.MURUGAN
Head , Dept of Computer Science
Shanmuga Insutries Arts & Science College ,Tamil Nadu , India.
profmuruga@gmail.com

Abstract

The importance of carrying out effective and sustainable agriculture is getting more and more obvious. In the past, additional fallow ground could be tilled to raise production. Nevertheless, even in industrialized countries agriculture can still improve on its overall yield. Modern technology, such as GPS-based tractors and sensor-aided fertilization, enables farmers to optimize their use of resources, economically and ecologically. However, these modern technologies create heaps of data that are not as easy to grasp and to evaluate as they have once been. Therefore, techniques or methods are required which use those data to their full capacity – clearly being a data mining task. This paper presents some experimental results on real agriculture data that aid in the first part of the data mining process: understanding and visualizing the data. We present interesting conclusions concerning fertilization strategies which result from data mining.

Key words: Precision Farming, Data Mining, Self-Organizing Maps, Neural Networks

Introduction

Recent worldwide economic development shows that agriculture will play a crucial role in sustaining economic growth, both in industrialized as well as in developing countries. In the latter countries agricultural development is still in its early stages and production improvements can easily be achieved by simple means like introduction of fertilization. In industrialized countries, on the other hand, even the agricultural sector is mostly quite industrialized itself, therefore improvements are harder to achieve. Nevertheless, due to the adoption of modern GPS technology and the use of ever more different sensors on the field, the term *precision farming* has been coined. According precision farming is the sampling, mapping, analysis and management of production areas that recognizes the spatial variability of the cropland.

In artificial intelligence terms, the area of precision farming (PF) is quite an interesting one as it involves methods and algorithms from numerous areas that the artificial intelligence community is familiar with. When analyzing the data flow that results from using PF, one is quickly reminded of *data mining*: an agriculturist collects data from his cropland (e.g., when fertilizing or harvesting) and would like to extract information from those data and use this information to his (economic) advantage. A simplified data flow model can be seen in 1. Therefore, it is clearly worthwhile to consider using AI techniques in the light of precision farming.

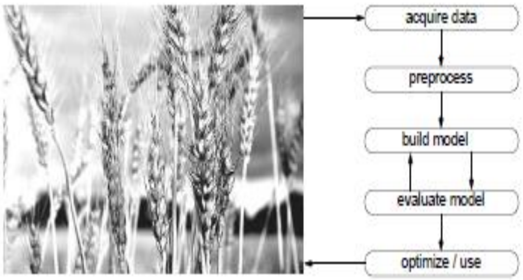

Fig. 1: Data mining for agriculture data

Data Description

The data available in this work have been obtained in the year 2006 on a field near K¨othen, north of Halle, Germany1 All information available for these 72- and 32-hectare fields2 was interpolated using kriging 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 field was subdivided into different strips, where various fertilization strategies were carried out. For an example of various managing strategies, this also shows the economic potential of PA technologies quite clearly. The field grew winter wheat, where nitrogen fertilizer was distributed over three application times during the growing season. Overall, there are seven input attributes – accompanied by the yield in 2006 as the target attribute. Those attributes will be described in the following. In total, for the smaller field (F131) there are 2278 records, for the larger field (F330) there are 4578 records, thereof none with missing values and none with outliers.

***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-383 are designed for agricultural use and can measure small-scale conductivity to a depth of about 1.5 meters. 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.

Fertilization Strategies

There were three different strategies that have been used to guide the nitrogen fertilization of the fields. F131 contains data resulting from two strategies (F, N) and F330 contains data from three strategies (F, N, S). The three strategies are as follows:

F – Uniform distribution of fertilizer according to long-term experience of the farmer

N – Fertilizer distribution was guided by an economic optimization with a multilayer perceptron model; the model was trained using the above data with the current year's yield as target variable that is to be predicted

S – Based on a special nitrogen sensor – the sensor's measurements are used to determine the amount of nitrogen fertilizer that is to be applied.

Using Multi-Layer Perceptions and Self-organizing Maps Approach

This section deals with the basic techniques that we used to model and visualize the agricultural yield data. For modeling, we have used Multi-Layer Perceptrons, as discussed. To visualize the data we will use Self-Organizing Maps (SOMs). Therefore, SOMs will comprise the main part of this section.

Multi-Layer Perceptrons for Modeling

In recent years, we have modeled the available data using a multi-layer perceptron (MLP). To gain more insights into what the MLP has learned, in this paper we will use self-organizing maps to try to better understand the data and the modeling process that underlies MLPs. In, neural networks have been used for optimization of fertilizer usage for wheat, in the process has been carried out for corn. In we could show that MLPs can be used for predicting current year's yield. For a detailed discussion of the used MLP structure and parameters, we basically used a feed forward-back propagation multi-layer perceptron with two hidden layers. The network parameters such as the hidden layer sizes were determined experimentally. A prediction accuracy of between 0.45 and 0.55 metric tons per hectare (100 × 100 metres) at an average yield of 9.14 *tha* could be achieved by using this modeling technique.

Self-Organizing Maps for Visualization

Our approach of using SOMs is motivated by the need to better understand the available yield data and extract knowledge from those data. SOMs have been shown to be a practical tool for data visualization. Moreover, SOMs can be used for prediction and correlation analysis, again, mostly visually. As such, the main focus in explaining Self-Organizing Maps in the following will be on the visual analysis of the resulting maps.

Self-Organizing Maps have been invented in the 1990s by Teuvo Kohonen. They are based on unsupervised competitive learning, which causes the training to be entirely data-driven and the neurons on the map to compete with each other. Supervised algorithms like MLPs or Support Vector Machines require the target attribute's values for each data vector to be known in advance whereas SOMs do not have this limitation.

Grid and Neigborhood: An important feature of SOMs that distinguishes them from Vector Quantisation techniques is that the neurons are organized on a regular grid. During training, not only the Best-Matching Neuron, but also its topological neighbors are updated. With those prerequisites, SOMs can be seen as a scaling method which projects data from a high-dimensional input space onto a typically two-dimensional map, preserving similarities between input vectors in the projection.

Structure: A SOM is formed of neurons located on a usually two-dimensional grid having a rectangular or hexagonal topology. Each neuron of the map is represented by a weight vector $mi = [mi1,min]T$, where n is equal to the respective dimension of the input vectors. The map's neurons are connected to adjacent neurons by a neighborhood relationship, superimposing the structure of the map. The number of neurons on the map determines the granularity of the resulting mapping, which, in turn, influences the accuracy and generalization capabilities of the SOM.

Training: After an initialization phase, the training phase begins. One sample vector **x** from the input data set is chosen and the similarity between the sample and each of the neurons on the map is calculated. The Best-Matching Unit (BMU) is determined: its weight vector is most similar to **x**. The weight vector of the BMU and its topological neighbors are updated, i.e. moved closer to the input vector.

The training is usually carried out in two phases: the first phase has relatively large learning rate and neighborhood radius values to help the map adapt towards new data. The second phase features smaller values for the learning rate and the radius to fine-tune the map.

Visualization: The reference vectors of the SOM can be visualized via component plane visualization. The trained SOM can be seen as multi-tiered with the components of the vectors describing horizontal layers themselves and the reference vectors being orthogonal to these layers. From the component planes the distribution of the component values and possible correlations between components can be obtained easily.

Experimental Results

This section will present some of the experimental results that we have obtained using SOMs on agricultural data. The first two parts will deal with the analysis of the maps generated from the complete data set (containing different fertilization strategies). The subsequent two parts will deal with those subsets of the data where a MLP has been used for yield prediction and optimization

Results for F131-all

The full F131-all dataset consists of the **F** and **N** fertilization strategies where each data record is labeled accordingly. After training the SOM using the preset heuristics from the toolbox [14], the labeled map those results. The corresponding U-Matrix that confirms the clear separability of the two fertilization strategies is shown in amount of fertilizer for the three different fertilization times is projected onto the same SOM. On those three maps it can also be seen that the different strategies are clearly separated on the maps. Another result can be seen. As should be expected, the REIP49 value (which is an indicator of current vegetation on the field) correlates with the YIELD06 attribute.

Results for F330-net

As in the preceding section, F330-net represents a subset of F330-all: it contains those data records from F330-all that were labeled *N*, i.e. in those field parts the MLP predictor was used for fertilizer optimization. Again, to convey a connection: the MLP has learned that where YIELD05 was high (lower left of map), there is less need of N1 fertilizer whereas the rest of the field needs a high amount. For N2, another network is trained with more input, now N2 and YIELD05 seem to correlate, although the correlation is not as clear as with the F131-net dataset. Furthermore, it is expected that REIP49 and YIELD06 correlate, even the EM38 value for electromagnetic conductivity correlates with the said attributes. Additionally, the corresponding scatter plot in shows a separation between clusters of low EM38/YIELD06 values and high EM38/YIELD06 values.

From the agricultural point of view, the F330 field is quite different from the one where the F131 data set was obtained; they are located 5.7km away from each other. This difference can be clearly shown on the SOMs. So, even though the fields are quite close, it is definitely necessary to have different small-scale and fine-granular fertilization and farming strategies.

5 Conclusions

In this paper we have presented a novel application of self-organizing maps by using them on agricultural yield data. After a thorough description and statistical analysis of the available data sets, we briefly outlined the advantages of self-organizing maps in data visualization. A hypothesis on the differences between two fields could clearly be confirmed by using SOMs. We presented further results, which are very promising and show that correlations and interdependencies in the data sets can easily be assessed by visual inspection of the resulting component planes of the self organizing map. Those results are of immediate practical usefulness and demonstrate the advantage of using data mining techniques in agriculture.

References

1. Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. WEBSOM—selforganizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, pages 310–315. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

2. T. Jensen, A. Apan, F. Young, and L. Zeller. Detecting the attributes of a wheat crop using
digital imagery acquired from a low-altitude platform. *Comput. Electron. Agric.*, 59(1-2):66–
77, 2007.

3. T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *Neural Networks, IEEE Transactions on*,11(3):574–585, 2000.

4. Teuvo Kohonen. *Self-Organizing Maps*. Springer, December 2000.

5. J. Liu, J. R. Miller, D. Haboudane, and E. Pattey. Exploring the relationship between red edge parameters and crop variables for precision agriculture. In *2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 2, pages 1276–1279 vol.2, 2004.

6. E. M. Middleton, P. K. E. Campbell, J. E. Mcmurtrey, L. A. Corp, L. M. Butcher, and E. W.
Chappelle. "Red edge" optical properties of corn leaves from different nitrogen regimes. In *2002 IEEE International Geoscience and Remote Sensing Symposium*, volume 4, pages 2208–2210 vol.4, 2002.

7. D. Pokrajac and Z. Obradovic. Neural network-based software for fertilizer optimization in precision farming. In *Int. Joint Conf. on Neural Networks 2001*, volume 3, pages 2110–2115, 2001.

8. Georg Ruß, Rudolf Kruse, Martin Schneider, and PeterWagner. Estimation of neural network parameters for wheat yield prediction. In *Proceedings of theWCC 2008*, Science and Business Media. Springer, 2008. (to appear).

9. Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Optimizing wheat yield prediction using different topologies of neural networks. In Jos´e Luis Verdegay, Manuel Ojeda-
Aciego, and Luis Magdalena, editors, *Proceedings of the International Conference*

*on InformationProcessing and Management of Uncertainty in Knowledge-Based Systems (IPMU-08),*pages 576–582. University of M´alaga, June 2008.

10. Georg Ruß, Rudolf Kruse, Peter Wagner, and Martin Schneider. Data mining with neural networks for wheat yield prediction. In *Advances in Data Mining*. Springer Verlag, 2008. (to appear).