Clustering-Regression-Ordering Steps for Knowledge Discovery in Spatial Databases*

Aleksandar Lazarevic[1], Xiaowei Xu[2], Tim Fiez[3] and Zoran Obradovic[1]
alazarev@eecs.wsu.edu, Xiaowei.Xu@mchp.siemens.de, tfiez@wsu.edu, zoran@eecs.wsu.edu
[1]School of Electrical Engineering and Computer Science and [3]Department of Crop and Soil Sciences
Washington State University, Pullman, WA 99164, USA
[2]Siemens AG, Corporate Technology, Information and Communications
Otto-Hahn-Ring 6, 81739 Munich, Germany

Abstract

Precision agriculture is a new approach to farming in which environmental characteristics at a sub-field level are used to guide crop production decisions. Instead of applying management actions and production inputs uniformly across entire fields, they are varied to match site-specific needs. A first step in this process is to define spatial regions having similar characteristics and to build local regression models describing the relationship between field characteristics and yield. From these yield prediction models, one can then determine optimum production input levels. Discovery of "similar" regions in fields is done by applying the DBSCAN clustering algorithm on data from more than one field ignoring spatial attributes (x and y coordinates) and the corresponding yield values. Using these models, constructed on training field regions of obtained clusters, we aim to achieve better prediction on identified regions than using global prediction models. The experimental results on real life agriculture data show observable improvements in prediction accuracy, although there are many unresolved issues in applying the proposed method in practice.

Purpose

Technological advances, such as global positioning systems, combine-mounted on-the-go yield monitors, and computer controlled variable rate application equipment, provide an opportunity for improving upon the traditional approach of treating agricultural fields as homogenous data distributions. In precision agriculture, environmental characteristics at a sub-field level are used to guide crop production decisions [6]. Instead of applying management actions and production inputs uniformly across entire fields, they are varied to better match site-specific needs thus increasing economic returns and improving environmental stewardship. Lower costs and new sensor technologies are enabling agriculture producers to collect large quantities of site-specific data from which future site-specific management decisions can be derived. However, methodologies to efficiently interpret the meaning of these large and multi-featured data sets are lacking. Site-specific variability in crop yield and optimum management actions can arise in two primary ways. First, levels of various driving variables almost always change throughout a field (e.g. weed density, soil N content, and soil depth). Second, the response to a given level of a driving variable can change within a field because of interactions with other driving variables. This results in poor site-specific recommendations when one uses a global recommendation equation that considers the variability of a single or a few driving variables [3,6].

The problem of yield prediction in agriculture is extremely complex since large number of attributes influence yield. In addition, there are differences in data distributions and significant amounts of noise can exist in data. Therefore, it appears necessary to develop local recommendation models that are adapted to specific subsets of the wide range of environments that can exist in fields even in a small geographic area. One recent approach towards such a modeling is to develop a sequence of local regressors each having a good fit on a particular training data subset. Distribution models are constructed for identified subsets, and are used to decide which regressor is most appropriate for each test data point [8].

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

A new approach for developing locally adapted models is explored in this paper. Given training and test fields, we first define more homogenous spatial regions in both fields. Spatial regions on the training field will have similar characteristics in attribute space to corresponding spatial regions in the test field. The next step is to build local regression models on spatial regions inside the training field, describing the relationship between field characteristics and yield. Using these models locally on corresponding spatial test field regions, we hope to achieve better prediction on identified regions than using global prediction models.

More precisely, this paper suggests clustering followed by local regression in order to identify site-specific yield prediction models from which optimum production input levels could be computed. This is followed by similarity-based competency ordering, which is used to identify the appropriate local regression model when making predictions for unseen fields.

Method

Given a rich feature set, partitioning a field into spatial regions having similar attributes (driving variables) should result in regions of similar yield response. Hence, the data from all fields were analyzed in order to define spatial regions having similar characteristics. Next, regression models were built to describe the relationship between attributes and yield on the training field subset of identified spatial regions.

To eliminate irrelevant and highly correlated features, regression-based feature selection was used for continuous target values and classification-based feature selection for discrete target values. The regression based feature selection process was performed through performance feedback forward selection and backward elimination search techniques based on linear regression mean square error (MSE) minimization. The classification based feature selection algorithms involved inter-class and probabilistic selection criteria using Euclidean and Mahalanobis distance, respectively [4]. In addition to sequential backward and forward search applied with both criteria, the branch and bound search was also used with Mahalanobis distance. To test feature stability, feature selection algorithms were applied to different data subsets, and the most stable features were selected.

In contrast to feature selection where a decision is target-based, variance-based dimensionality reduction through feature extraction is also considered. Here, linear Principal Components Analysis [4] and non-linear dimensionality reduction using 4-layer feedforward neural networks (NN) [1] was employed. The targets used to train these NNs were the input vectors themselves, so that the network is attempting to map each input vector onto itself. We can view this NN as two successive functional mappings. The first mapping, defined by the first two layers, projects the original d-dimensional data into a r-dimensional sub-space ($r < d$) defined by the activations of the units in the second hidden layer with r neurons. Similarly, the last two layers of the NN define an inverse functional mapping from the r-dimensional sub-space back into the original d-dimensional space.

Using the features derived through the feature selection and extraction procedures, the DBSCAN clustering algorithm [2,9] was used to partition fields into "similar" regions ignoring the spatial attributes (x and y coordinates) and the yield value. The DBSCAN algorithm was applied to merged training and testing field data. These fields need not be adjacent as the x and y coordinates were ignored in the clustering process. The DBSCAN algorithm relies on a density-based notion of clusters and was designed to discover clusters of arbitrary shape efficiently. The key idea of a density-based cluster is that for each point of a cluster its *Eps*-neighborhood for some given *Eps* > 0 has to contain at least a minimum number of points (*MinPts*), (i.e. the density in the *Eps*-neighborhood of points has to exceed some threshold). Furthermore, the typical density of points inside clusters is considerably higher than outside of clusters. DBSCAN uses a simple but effective heuristic for determining the parameters *Eps* and *MinPts* of the smallest cluster in the database.

We used the DBSCAN algorithm in an unsupervised manner, using different variations of attributes obtained through feature selection and extraction or through normalization of the original data. As a result, we have the partitions P_i, which are generally spread in both training and test field parts. Since the resulting partitions P_i are constructed without considering spatial information, the next step is to identify the largest contiguous clusters C_i inside the training part of partitions P_i, and also the largest contiguous clusters T_i inside the test field part of partitions P_i (Figure 1). The identification of C_i and T_i is performed by collecting all the neighboring (x, y) points belonging to P_i. Note there may be 2 or more such regions in the fields.

To further specialize prediction, in each training cluster C_i we also identified subsets L_i, A_i and H_i by assigning C_i data into three equal-size parts according to the yield. Hence, the subset L_i corresponds to the lowest 33% of the yield in C_i (Figure 2) while subsets A_i and H_i represent the average 33% and the highest 33% of the yield in cluster C_i.

For each C_i, L_i, A_i and H_i we ordered corresponding test-field data (T_i) in the P_i according to their distance from the T_i, L_i, A_i and H_i center points determined by mean (see Figure 2). This is measured based on Euclidean or Mahalanobis distance among the various subsets of

attributes obtained through the preprocessing steps. Due to possible feature instability, we performed an independent feature selection process for each cluster C_i and used region-specific features for computing distance.

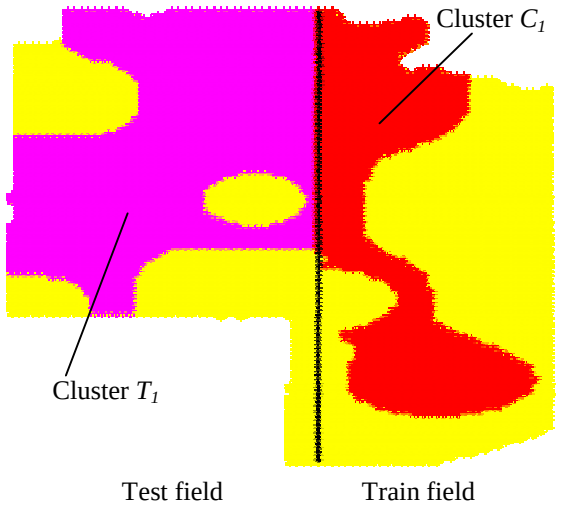

Figure 1. A spatial cluster P_1, obtained by DBSCAN on merged fields, is split into a training cluster C_1 and a corresponding test cluster T_1

An alternative to ordering the test field data, the weighted majority k-Nearest Neighbor algorithm with weights inversely proportional to the distances from the center point was also considered [4].

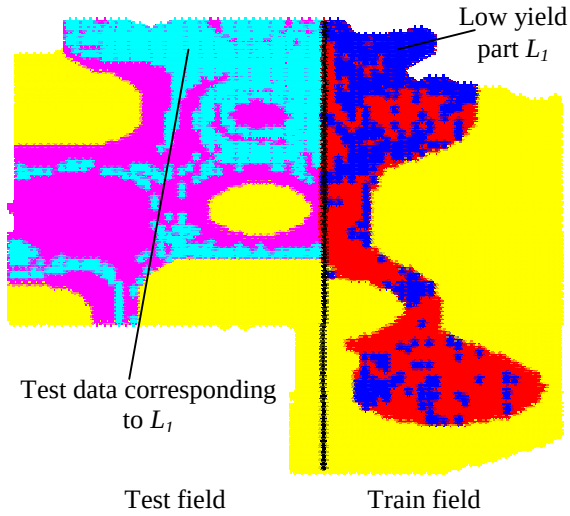

Figure 2. Low yield subset L_1 of training cluster C_1 and corresponding test data identified through Euclidean distance-based ordering

Finally, linear regression models and multilayer (2-layered) feedforward neural network (NN) regression models, with back-propagation learning [11], were trained on each spatial part C_i, L_i, A_i and H_i, and were applied to the corresponding neighborhood parts in the test field. For each of these models, we measured the Mean Square Error (MSE) of yield prediction on identified test parts.

Results

Our experiments were performed using spatial data from a 280 ha wheat field located in southern Idaho [7]. All features were interpolated to a 10x10 m grid resulting in 7036 patterns. The data set contained 17 soil attributes and the corresponding crop yield. The soil attributes included levels of boron, calcium, copper, iron, magnesium, manganese, nitrate nitrogen, potassium, sodium, organic matter, organic nitrogen, phosphorus, salts, sulfur, and zinc and soil cation exchange capacity (CEC) and soil pH.

The feature selection processes were used to select the 2, 3, and 4 most relevant features (Table 1).

Number of features	List of selected features
2	CEC, Iron
3	CEC, Iron, Manganese
4	CEC, Iron, Manganese, Salts

Table 1. The selected features

The PCA and non-linear Feature Extraction procedures were also used to reduce data dimensionality. Here, projections to 2 dimensions explained most of the variance and there was little improvement from additional dimensions.

A typical way to test generalization capabilities of regression models is to split the data into a training and a test set at random. However, for spatial domains such an approach is likely to results in overly optimistic estimates of prediction error [10]. Therefore, the test field was spatially separated from the training field such that both were of an equal area, as shown in Figure 1.

We clustered the train and test patterns using each of the feature subsets obtained through the feature selection procedures. Changing the *Eps* and *MinPts* parameters of DBSCAN algorithm changed the size of resulting clusters. By increasing the *MinPts* parameter, which generally means increasing the *Eps* parameter too, we can reduce the number of resulting clusters. The results shown in Table 2 represent clusters obtained with *Eps* and *MinPts* values which minimized the number of clusters and the largest cluster size. The best clustering result as measured by relative cluster sizes and number of clusters was obtained using the 2 feature data set containing only CEC and iron

levels (Table 2). Using 3 or 4 features for the clustering process usually resulted in a huge first cluster and a number of small clusters. This was even more pronounced when clustering with more than 4 features.

Numbers of used features	2	3	4
Number of clusters	6	7	7
First cluster size (percentage of the entire field)	44%	69%	75%

Table 2. Obtained clusters

Applying the DBSCAN algorithm using features obtained through PCA and non-linear feature extraction resulted in an even larger first cluster and many small ones. For example, when clustering data projected to 2 dimensions, the minimal size of the largest cluster was 77% of the entire field, and the total number of clusters was 11.

Therefore, all reported prediction modeling was performed on the 6 clusters identified using only 2 features (CEC and iron levels). Prediction models were developed for the entire training field, each cluster in the training field, and each part, L_i, A_i, and H_i, of each cluster in the training field. Test field data were ordered by Euclidean and Mahalanobis distance to determine the appropriate model to use to predict wheat yield.

We developed neural network models with 2, 3 and 4 features, and it turned out that neural network models trained with only 2 features gave the best prediction results. The MSE for these models applied to the largest cluster and using Euclidean distance to determine the appropriate model is shown in Table 3. Results for the other clusters are comparable to those for the largest cluster.

Models trained with 2 features		MSE on test	
Trained on	Tested on	NN model	Linear model
Entire train field	Entire test field	458	397
Entire train field	Cluster T_1	389	395
Cluster C_1		354	341
Entire train field	Test data corresponding to L_1	472	405
Cluster C_1		491	351
Low yield part L_1		318	388
Entire train field	Test data corresponding to A_1	415	399
Cluster C_1		404	346
Average yield part A_1		387	390
Entire train field	Test data corresponding to H_1	478	382
Cluster C_1		461	416
High yield part H_1		943	964

Table 3. MSE on site specific regions identified using Euclidean distance

Analyzing the data from Table 3, the method of building local site-specific regression models outperformed the global models for the low and the average yield fractions of the largest cluster. It can be noticed, that a model built on the entire training field and tested on cluster T_1 does not have a MSE equal to the average of the MSE, when the same model is applied to test data corresponding to L_1, A_1 and H_1. This phenomenon is due to the overlapping of identified test data parts, which when aggregated do not form the cluster T_i. The linear model resulted in a lower MSE than that of the NN for the average yield partition, A_1, of the cluster apparently because of the small yield variance in that partition. The large MSE of site specific models (almost twice of that for global models, last part of Table 3) observed for the high yield cluster component H_i indicates a poor mapping of test fields points to the appropriate model. The training results for the H_1 data were similar to that for the L_1 and A_1 data. Thus, the Euclidean distance metric does not appear to be a good measure of similarity between test points and high yield training points.

Models trained with 2 features		MSE on test	
Trained on	Tested on	NN model	Linear model
Entire train field	Test data corresponding to L_1	445	356
Cluster C_1		438	375
Low yield part L_1		278	348
Entire train field	Test data corresponding to A_1	367	332
Cluster C_1		365	353
Average yield part A_1		363	342
Entire train field	Test data corresponding to H_1	417	353
Cluster C_1		541	394
High yield part H_1		808	819

Table 4. MSE on regions identified using Mahalanobis distance

The MSE resulting from using Mahalanobis distance instead of Euclidean distance to determine the appropriate model is shown in Table 4. The use of Mahalanobis distances has shown an observable improvement in prediction accuracy in each part, but the MSE on the test part corresponding to H_1 is still unacceptably high when compared to the global model. In a final attempt to find a better distance measurement for the high yield patterns, we also tried a weighted majority k-Nearest Neighbor approach. However, results were even worse than using the Euclidean or Mahalanobis distances. To overcome overlapping problem, additional distance metrics may need to be examined like Bhatacharya [4] and the distance metric proposed by Hand and Henley [5].

To further examine our similarity metrics and to assess cluster homogeneity, we ordered the points within a single

cluster in test regions T_i, according to the Euclidean or Mahalanobis distance from the center of the corresponding training cluster C_i. It might be expected that errors for test points close to the training cluster center would be less than that for more distant test points. Error data from one of clusters is shown in Figure 3. The test-set data for Figure 3 were sorted according to their Euclidean distances from the center of the training cluster, and split into 6 equal size groups, from the nearest group 1 to the most distant group 6. This experiment clearly demonstrates that even better prediction may be possible with further site-specific localization.

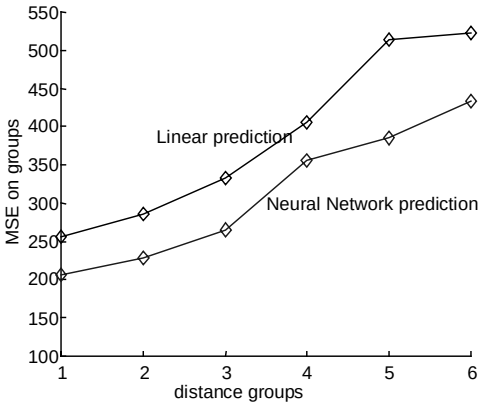

Figure 3. Test error analysis within a single test cluster

New aspects of work

A new method for analyzing field information in precision agriculture is proposed. A sequence of non-spatial and spatial clustering steps followed by local regression modeling combined with ordering-based competency identification is proposed for spatial knowledge discovery. The new approach is successfully applied to precision agriculture management.

Conclusions

Results from actual precision agriculture data indicate that the process of density-based clustering followed by cluster specific model development can result in better yield prediction than a single global model in some cases. Further work is needed to refine the clustering process such that it does not digress to identifying one large cluster with most of the data and many small clusters containing the rest of the data. To overcome this problem, we are currently exploring a hierarchical clustering (i.e. repeatedly re-cluster the largest cluster until satisfactory cluster sizes are obtained).

Additionally, the distance metrics used to identify which model should be used for a test pattern were not uniformly adequate across low, average, and high yield data. The failure to properly associate patterns to high yield models needs to be addressed.

All the prediction models applied in this project exhibited a limited capability for yield prediction and there was little benefit from using neural networks over simple linear models. However, these results are probably due to the lack of appropriate driving variables to explain the yield variability.

References

[1] Bishop, C., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[2] Ester M., Kriegel H.-P., Sander J., Xu X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226-231, Portland, OR, 1996.

[3] Fiez, T., Miller, B.C., Pan, W.L.: "Assessment of Spatially Variable Nitrogen Fertilizer Management in Winter Wheat," *Journal of Production Agriculture*, Vol. 7, No. 1, pp. 86-93, 1994.

[4] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, CA, 1990.

[5] Hand, D.J., and Henley, W. E.: "Statistical Classification Methods in Consumer Credit Scoring," *Journal of Statistical Society*, Part 3., pp. 523-541, 1996.

[6] Hergert, G, Pan, W., Huggins, D., Grove, J., Peck, T., "Adequacy of Current Fertilizer Recommendation for Site-Specific Management", In Pierce F., "The state of Site-Specific Management for Agriculture," *American Society for agronomy, Crop Science Society of America*, Soil Science Society of America, chapter 13, pp. 283-300, 1997.

[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. Conf. on Geospatial Inf. in Agric. and Forestry*, Vol. I, Lake Buena Vista, Florida, pp. 206-210, June 1-3, 1998.

[8] Pokrajac, D., Fiez, T., Obradovic, D., Kwek, S. and Obradovic, Z.: "Distribution Comparison for Site-Specific Regression Modeling in Agriculture," *Proc. IEEE/INNS Int'l Conf. on Neural Neural Networks*, Washington, D.C., July, 1999, in press.

[9] Sander J., Ester M., Kriegel H.-P., Xu X.: "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications," *Data Mining and Knowledge Discovery, An International Journal*, Kluwer Academic Publishers, Vol. 2, No. 2, pp. 169-194, 1998.

[10] Vucetic, S., Fiez, T. and Obradovic, Z. "A Data Partitioning Scheme for Spatial Regression," *Proc. IEEE/INNS Int'l Conf. on Neural Neural Networks*, Washington, D.C., July 1999., in press.

[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.