Application of statistical machine learning algorithms in precision agriculture

Mohan Sridharan*[1], Prasanna Gowda[2]
1 Department of Electrical and Computer Engineering, The University of Auckland, NZ
2 Forage and Livestock Production Research Unit, US Department of Agriculture, USA

Abstract

Remote sensing can facilitate rapid collection of data in agriculture at relatively low cost. Advancements in unmanned aerial vehicles and sensor technology, along with a significant reduction in the cost of acquiring data, have enabled us to collect and process remote sensing data in real time. Approaches based on remote sensing data are widely used in precision agriculture for estimating crop and soil characteristics such as leaf area index, biomass, crop stress, evapotranspiration, crop yield, and soil organic matter. These approaches typically use predictive models (e.g., linear, quadratic, power or exponential) that are based on ordinary least square (OLS) regression. However, the performance of these predictive models deteriorates when the effects of sun-surface sensor geometry, background reflectance and atmosphere-induced variations on spectral reflectance or spectral vegetation indices are larger than the variations in the crop or soil characteristics of interest. Any errors in the predicted soil and crop characteristics may, in turn, adversely affect farm inputs, farm outputs and thus the net profits. In recent years, machine learning algorithms such as artificial neural networks, support vector machines and Gaussian processes are being explored for developing predictive models for agricultural applications, especially since these algorithms are known to provide more accurate predictions than OLS. In this paper, we describe and experimentally compare the accuracy of OLS and statistical machine learning models for estimating crop water use (or evapotranspiration). We show that models based on machine learning algorithms provide significant improvement in accuracy in comparison with a state of the art energy balance model based on OLS. We use this example to highlight the potential benefits of the use of statistical machine learning algorithms in precision agriculture.

Background

Evapotranspiration (ET) is the most important hydrologic process that plays an essential role in determining exchanges of energy and mass between the hydrosphere, atmosphere and biosphere (Sellers et al., 1996). In agriculture, accurate estimates of ET, which includes water evaporation from land and water surfaces, and transpiration by vegetation, is needed in any effort to improve water use efficiency. Daily grass or alfalfa reference ET values and crop coefficients are widely used to estimate the water demand of particular crops. ET varies spatially and seasonally according to weather conditions. Understanding these variations is essential for managers who are responsible for planning the management of water resources, especially in arid and semi-arid regions of the world where inadequate precipitation to meet crop water demand is supplemented by irrigation from surface and/or groundwater resources for sustainable crop production.

Conventional techniques for measuring ET fluxes at field scales require homogeneous surfaces and do not provide spatial trends, especially in regions with advective climatic conditions (Gowda et al., 2008). Land surface energy balance (EB) models, which use remote sensing data from ground to airborne to satellite platforms at different spatial resolutions, have been found to be promising to provide spatially distributed daily and seasonal reference ET at both field and regional scales. These models convert satellite sensed radiances into land surface characteristics such as albedo, leaf area index, vegetation indices, surface emissivity and surface temperature to estimate ET as a residual of the land surface energy balance equation. Many remote sensing algorithms are available for

estimating magnitude and trends in ET, e.g, Surface Energy Balance Algorithm for Land (SEBAL) by Bastiaanssen et al. (1998), Surface Energy Balance System (SEBS) by Su (2002), and Simplified Surface Energy Balance by Gowda et al. (2009). A detailed review of ET algorithms is presented in Gowda et al. (2008). These EB models are very complex to use and require advanced knowledge of remote sensing and agri-meteorology. Also, ET estimation errors seem to vary randomly between geographic regions, and may be due to errors in the estimation of crop and surface characteristics using empirical models that are sensitive to crop structure, soil background, landscape position and geographic location.

Many of the EB models and other statistical methods used to estimate reference ET (and other agricultural indicators) are often based on ordinary least square (OLS) regression methods. In addition, many of these models depend on accurate measurement of relevant parameters (e.g., weather parameters), which might require the use of equipment that is costly to obtain or maintain. In recent years, machine learning algorithms such as artificial neural network (ANN), support vector machines (SVM), and Gaussian Process (GP) have been shown to provide significantly better performance than models built using OLS methods. These algorithms have different advantages and limitations, and their suitability for a particular application depends on factors such as computational cost, availability of training samples etc. For instance, ANNs are capable of representing complex, nonlinear relationships in a variety of fields, including geology and hydrology (ASCE, 2000), but can result in local minima, over-fitting, and high computational cost in high dimensions. SVMs use basis functions to map data to other (potentially infinite) dimensions and construct an optimal separating hyperplane in the transformed space. SVMs have a simple geometric interpretation, avoid overfitting, find global solutions and use structural risk minimization methods (Vapnik, 1995). GPs explore the infinite space of functions to accurately capture the relationship between inputs and outputs---they have been used in different domains as sensor networks and climate informatics (Higdon, Lee & Holloman, 2003; Krause, Singh & Guestrin, 2008). Our prior work was one of the few existing instances to use GP for irrigation scheduling
(Holman et al., 2014).

To promote the widespread use of statistical machine learning algorithms for accurate estimation of agricultural indicators, we summarize below the result of developing ANN, SVM and GP-based models for estimating reference ET from Landsat Thematic data. We used lysimetric data (as the target output) to train these models, and compared the performance of these models with the documented performance of SEBS, an EB model used widely around the world.

Data and Methods

This study was conducted in the area covered by Landsat 5's path/row of 31/36 in the Southern High Plains (parts of the Texas High Plains and northeastern New Mexico) in south-central USA. The climate in this area is semiarid with highly variable rainfall---475 mm annual average with 348 mm during the summer growing season. The dominant soil in the study area is classified as a Pullman clay loam with low permeability. The major crops are corn, sorghum, winter wheat, and cotton, and the cropping season starts in May and ends in October. In this study, we used data collected over a period of four years.

The area considered in this study corresponded to four fields, and different crops were cultivated in these fields. As documented by Gowda et al. (2013), the current state of the art approach for estimating reference ET in these fields is SEBS. The accuracy of SEBS has been evaluated using soil water mass change-based daily ET values from four large monolithic precision weighing lysimeters located at the USDA-ARS Conservation and Production Research Laboratory (CPRL) in Bushland, USA. CPRL is in the northeastern corner of the Landsat scene and its geographic coordinates are 35o11'N, 102o 06'W, and elevation is 1170 m above mean sea level. Each lysimeter (3m x 3m x 2.4m in length, width and depth) is in the middle of a 4.7-ha field. Changes in lysimeter mass were measured using a data logger for determining ground truth ET. Daily ET was calculated as the difference between lysimeter mass recorded at 2330 of one day and 2330 of the next day to

determine mass losses (evaporation and transpiration) to which lysimeter mass gains (irrigation or precipitation) were added. Dryland cropping systems were managed on two lysimeter fields in the west and irrigated cropping systems are managed on two lysimeter fields in the east with a 10-span lateral-move sprinkler system. A grass reference ET weather station field (uniformly cut, 120-mm tall fescue blend grass on 0.31-ha area) is located adjacent to the eastern side of the irrigated lysimeter fields. Grass is the preferred reference crop in arid and semi-arid areas such as the Texas High Plains because it is more representative of reference conditions.

The data used in this study consisted of daily measurements of weather parameters, indices computed from remote sensing data, and lysimeter-based ET measurements, from each of the four fields in the study area. Specifically, we considered seven weather parameters: air temperature, wind speed, relative humidity, pressure, solar radiation, surface temperature, and net radiation. We also considered remote sensing data in the form of reflectance encoded in bands 1-7 in the Landsat images, to compute two indices: Normalized Difference Vegetation Index (NDVI), and Enhanced Vegetation Index (EVI). In addition, we consider the reference ET measurements obtained from the lysimeters in the study area as the ground truth, i.e., the target output values of our estimation models. We then formulated reference ET estimation as a supervised learning problem. Specifically, training data consisted of N vectors of inputs $X= \{x1,..., xN\}$ and target outputs $Y= \{y1,...,yN\}$, and test data consisted of previously unseen inputs and target outputs. The reference ET estimation models constructed using the training data had to process input vectors of the test dataset to provide estimates of reference ET---these outputs were then compared with the ground truth target outputs. The input vector of the training dataset and test dataset can consist of some combination of weather parameter measurements and indices computed from the Landsat images of each of the four fields. We experimentally observed that the most accurate ET estimates were obtained when all seven weather parameters were included in input vector. We thus considered three types of input vectors (1) seven weather parameters; (2) seven weather parameters with NDVI; and (3) seven weather parameters with EVI. The output value for each input vector was the corresponding reference ET measurement provided by the lysimeter in the corresponding field.

The reference ET estimation models are evaluated using three performance measures: coefficient of determination (R2), root mean square error (RMSE), and Nash-Sutcliffe efficiency (NSE). The value of R2 describes the proportion of variability in the observed data that is explained by the model. R2 ranges from 0 to 1, with a higher value indicating a better goodness of fit. For instance, R2=1 with an intercept of 0 and slope of 1 indicates a perfect fit between the observed and modeled data. An RMSE of 0 also indicates a perfect fit. RMSE is usually computed as a percentage of the observed mean,
e.g., RMSE < 50% is usually considered low. The RMSE measure is appropriate for our study because it provides values in the same units as the values that are to be estimated (i.e., reference ET). The NSE is a common efficiency measure that compares variance in estimations with the variance in the measured data. NSE ranges between negative infinity and 1; NSE values closer to 1 are more accurate and an NSE of 1 represents an optimal model. Negative NSE values indicate that the mean of the observations is more accurate than the model estimation. More information on these and other such performance statistics can be found in Moriasi et al. (2007).

As stated earlier, we adapted three popular machine learning algorithms for modeling the complex, nonlinear relationships that need to be considered to estimate reference ET accurately: Artificial Neural Network (ANN). Support Vector Machine (SVM), and Gaussian Process (GP). We hypothesized that the performance of these models would be better than (or at least as good as) the reported performance of a state of the art EB model (Gowda et al., 2013).

Implementation and Results

The software implementation used in our experimental trials was based on the WEKA open source machine learning library (Hall et al., 2009). WEKA includes Java implementation of popular machine learning algorithms such as linear regression, ANN, SVM, and GP. The library also has evaluation schemes that can be used to compare the performance of different algorithms over different datasets.

We adapted the existing implementations to fit our needs. The individual algorithms also have some parameters that can be tuned to improve performance. We experimentally tuned the value of these parameters. We then performed two types of experiments. First, we took the entire dataset and performed a standard 10-fold cross-validation (also called “leave one out”) analysis. Second, we split the available data into a training set (70% of data) and test set (30% of data) by randomly selecting samples from the available data---the machine learning models were built using the training set and evaluated on the test set. As stated earlier, three different types of input vectors were considered: seven weather parameters, seven weather parameters with NDVI, and seven weather parameters with EVI. The target output (in all cases) was the reference ET measurements from the lysimeters. The measures used for comparing performance were R2, RMSE, and NSE.

As an illustrative example, Table 1 summarizes the results obtained with 10-fold cross-validation. The results indicate that the best performance is provided by the SVM-based model (with an RBF kernel) for ET estimation. We also observe that performance is similar with either NDVI or EVI included in the input along with the seven weather parameters. Also, including either NDVI or EVI improves performance in comparison with only using the seven weather parameters. A similar performance was obtained when a 70%-30% split of data was used for training and testing the statistical machine learning models---SVM-based model provides the best results, and there is an improvement in performance when either NDVI or EVI is included in the input data of weather parameters. The best performance obtained with the machine learning models is also better than the performance of the SEBS model reported by Gowda et al. (2013).

Table 1. Performance measures R2, RMSE, and NSE (after ten-fold cross-validation) for models estimating reference ET from weather attributes and indices---SVMs provide the best performance.

Model	**7 weather parameters**			**7 weather parameters+NDVI**			**7 weather attributes+EVI**		
	RMSE	R^2	NSE	RMSE	R^2	NSE	RMSE	R^2	NSE
LR	0.15	0.70	0.70	0.09	0.88	0.88	0.10	0.86	0.86
ANN	0.11	0.82	0.81	0.10	0.86	0.84	0.11	0.85	0.83
SVM (RBF)	**0.08**	**0.91**	**0.91**	**0.07**	**0.93**	**0.93**	**0.07**	**0.93**	**0.93**
GP (RBF)	0.10	0.88	0.85	0.10	0.89	0.84	0.10	0.90	0.84

Conclusion and discussion

Accurate estimation of reference ET is essential for optimal crop water use. Existing state of the art energy balance (EB)-based models and other statistical methods used to estimate reference ET are often based on ordinary least square (OLS) regression methods. In addition, many of these models depend on accurate measurement of relevant parameters (e.g., weather parameters), which may require the use of equipment that is costly to obtain or maintain. In this paper, we explored the use of statistical machine learning algorithms for the underlying regression task. Results obtained on data from a region in West Texas (USA) indicate that the machine learning-based models provide a more accurate estimate of reference ET at a fraction of the computational complexity and cost associated with the EB-based models.

These results open up multiple directions of further research. For instance, machine learning-based models built from historical data can be used to accurately estimate other agricultural indicators such as ground cover, leaf area index, chlorophyll level etc. The current state of the art methods used for such problems in agriculture have limitations similar to those of using EB-based models for reference ET estimation. In addition, such models can be built for other estimation and prediction problems in precision agriculture, e.g., crop yield prediction and gross primary productivity from remote sensing data. In the long-term, machine learning-based models show promise for automating and significantly improving accuracy, computational efficiency and cost of various tasks in precision agriculture.

Acknowledgments

The authors thank Ning Wu and Jing Lu, who helped generate the results reported in this paper. This work was supported in part by the University of Auckland Research Foundation's Distinguished Visitor Award.

USDA EEO Disclaimer

References

ASCE Task Committee on applications of artificial neural networks in hydrology (2000). Artificial Neural Networks in Hydrology I: Preliminary Concepts. Journal of Hydrol. Eng. 5: 115–123.

Bastiaanssen WG, Menenti M, Feddes RA, Holtslag AA 1998. A remote sensing surface energy balance algorithm for land (SEBAL): Formulation. Journal of Hydrology 212–213, 198–212.

Gowda PH, Howell TA, Paul G, Marek TH, Su B, Copeland KS 2013. Deriving hourly evapotranspiration with SEBS: A lysimetric evaluation. Vadoze Zone Journal 12: 1–11.

Gowda PH, Chavez JL, Colaizzi PD, Evett SR, Howell TA, Tolk JA 2008. ET mapping for agricultural water management: Present status and challenges. Irrigation Science 26: 223–237.

Gowda PH, Senay G, Howell TA, Marek TH 2009. Lysimetric evaluation of simplified energy balance approach in the Texas High Plains. Applied Engineering in Agriculture 25: 665–669.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 p.

Higdon K, Lee H, Holloman C 2003. Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. *In:* Bayesian Statistics, Oxford University Press.

Holman D, Sridharan M, Gowda P, Porter D, Marek T, Howell T, Moorhead J 2014. Gaussian process models for reference ET estimation from alternative meteorological data sources. Journal of Hydrology 517: 28–35.

Krause A, Singh A, Guestrin C 2008. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. Journal of Machine Learning Research 9: 235–284.

Moriasi DN, Arnold JG, Liew MWV, Bingner RL, Harmel RD, Veith TL 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE
50: 885–900.

Sellers PJ, Randall DA, Collatz GJ, Berry JA, Field CB, Dazlich DA, Zhang C, Collelo GD, Nounoua L 1996. A revised land surface parameterization (SiB2) for atmospheric GCMS, Part 1: Model formulation. Journal of Climate 9: 676–705.

Su B 2002. The surface energy balance system (SEBS) for estimation of turbulent fluxes. Hydrology and Earth Systems Science 6: 85–99.

Vapnik VN 1995. The nature of statistical learning theory. Springer.