Considerations about Application of Machine Learning to the Prediction of Sigatoka Disease

Luis-Alexander Calvo-Valverde[1], Mauricio Guzmán-Quesada[2]

and José-Antonio Guzmán-Álvarez[3]

[1] Centro de Investigaciones en Computación
Instituto Tecnológico de Costa Rica
DOCINADE
Cartago, Costa Rica

[2,3] Dirección de Investigaciones
Corporación Bananera Nacional S.A.
Guápiles, Costa Rica

lcalvo@itcr.ac.cr, mguzman@corbana.co.cr, jguzman@corbana.co.cr

ABSTRACT

One issue that has occupied the attention of humanity is the production of food and view it from several perspectives: the quality of the seed, the production process, diseases that affect productivity, the effect of climate and location, in example. As a contribution to the above situation, this paper presents the application of one discipline of artificial intelligence, known as machine learning, which involves the study of computer algorithms that improve automatically through experience. This type of learning has been used in applications ranging from data mining to discover rules in large datasets, to information filtering systems that automatically learn user interests. As a particular case, the Corporación Bananera Nacional of Costa Rica (Corbana) has stations measuring meteorological variables. These variables measured are temperature, precipitation, humidity, wind speed, among others. Corbana is interested in relating this information with the spread of a disease that affects production; this disease is the sigatoka. In addition, this organization recorded weekly in several of his research areas the following variables related to that disease: state of evolution, severity leaf 2, severity leaf 3, among others. With this data and using machine learning algorithms; they want to make predictions. This work presents the results of applying various machine learning algorithms to available data, in example, artificial neural network, support vector machines regression. These first conclusions will be tuned in the future

Keywords: Machine learning, artificial neural network, support vector regression, disease prediction, banana, black Sigatoka, Costa Rica.

1 INTRODUCTION

The Food and Agriculture Organization of United Nations (2012) in its annual report indicates that the main conclusion of the assessment carried out worldwide is that agriculture appears to be driven by an expansion demand being covered mostly by new and emerging exporters, rather than

L. Calvo, M. Guzmán and J. Guzmán. "Considerations about Application of Machine Learning to the Prediction of Sigatoka Disease". World Conference on Computers in Agriculture and Natural Resources, University of Costa Rica, San Jose Costa Rica, July 27th-30th, 2014. http://CIGRProceedings.org

traditional providers. However, the increase in the price of inputs and the cost of access from remote areas has led to increases food prices in real terms. The question is whether the production will be able to grow with demand in the coming years so that they achieve their real stabilize prices historical patterns, or if these prices continue to rise by the growing pressure of demand.

Nowadays there are efforts to apply machine learning methods for decision-making in agriculture, including the control of crop diseases. By example, Camargo et al. (2012) present an intelligent systems for the assessment of crop disorders, Huang et al. (2010) present a plant virus identification based on neural networks with evolutionary preprocessing, Kim et al. (2014) present a survey about crop pests prediction method using regression and machine learning technology and Zhao et al. (2013) present an intelligent agricultural forecasting system based on wireless sensor.

2 CONCEPTS

2.1 Machine Learning (ML)

It refers to the study of computer algorithms that improve automatically through experience. This type of learning has been used in applications ranging from data mining to discover rules in large datasets, to information filtering systems that automatically learn user interests. Murphy (2012) says that machine learning is a set of methods that can automatically detect patterns in the data, and then use the discovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (e.g. to plain how to collect more data).

2.2 Black Sigatoka Disease

Black Sigatoka, disease caused by the fungus *Mycosphaerella fijiensis Morelet*, is the main problem phytopathologic of banana and plantain crops in Central America (Marin Vargas & Romero Calderón, 1995).

This disease attacks the leaves of plants producing a rapid deterioration of the leaf area. It affects the growth and productivity of plants by decreasing photosynthetic capacity. Also causes a reduction in quality of the fruit (Marin Vargas & Romero Calderón, 1995).

The climate has a major effect on the behavior of the black Sigatoka. Precipitation, temperature, relative humidity and wind are the main climatic variables affecting the development of this disease (Marin Vargas & Romero Calderón, 1995).

2.3 Biological Warning System

This system measures the state of development of the disease, to determine the time in which apply fungicides (Marin Vargas & Romero Calderón, 1995).

This system is based on two components, a climate component, that is given by the *Piche* evaporation and other biological, given by the state of progress or the rate of disease development. Originally, this system was designed to work with young plants. The selected plant should have normal growth, and be in a suitable site, allowing behavioral considered as representative of the estate. It requires that the plant begins with 5 or 6 true leaves (Marin Vargas & Romero Calderón, 1995).

2.4 Artificial Neural Network (ANN)

An artificial neural network is a directed graph with the following properties (Martín del Brío & Sanz Molina, 2010): 1) to each node *i* a state variable x_i is associated, 2) to each connection *(i, j)* of the node *i* and *j*, a weight $w_{ij} \epsilon R$ is associated, 3) to each node *i* is associated a threshold Θ_i, 4) to each node *i* a $f_i(x_j, w_{ij}, \Theta_i)$ function is defined, which depends on the weights of their connections, of the threshold and the states of the node *j* to him connected. This function provides the new status of the node.

The nodes are the neurons and the connections are the synapsis. There are son important concepts: 1) input neuron is a neuron that without input synapsis, 2) output neuron is a neuron without output synapsis, 3) neurons that are neither input nor output are called hidden neurons, 4) a network is unidirectional if do not have loops, 5) a network is recurrent when the flow of information can be found a loop back and forth.

2.5 Echo State Networks (ESN)

Recurrent Neural Networks (RNN) -these have directed cycles in their connection graph- are useful for temporal patterns, but when they are trained with backpropagation method, they are very slow. Echo State Network is an alternative training method to solve that problem. This ESN is based on the observation that if a random RNN possesses certain algebraic properties, training only a linear readout from it is often sufficient to achieve excellent performance in practical applications. The untrained RNN part of an ESN is called a dynamical reservoir, and the resulting states *x(n)* are termed echoes of its input history (Lukoševičius & Jaeger, 2009).

Given a dataset with a number of *K* variables which varies depending on a *n* variable, it want the creation of a first layer of neurons *u* with *K* elements to process input data (Lukoševičius M. , 2012).. After, a second layer of neurons with *N* elements is generated, called *reservoir* (represented mathematically with the variable *X*), which receives input data and processes them in parallel and recurrent form. Finally, these signals are sent to an output layer *y* with *L* output variables as shown in the following Figure.

.

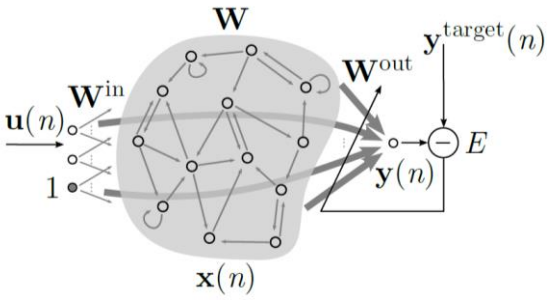

Figure 1: An echo state network (Lukosevicius, 2012)

The connections between the different elements of an Echo State Network have weights randomly generated. The weights of the internal connections of the reservoir (*W*) as well as the weights of the input layer (W_{in}), after being generated are set statically during all stages of implementation of

the algorithm. The weights between the reservoir and the output layer (W_{out}) are subject to changes of a supervised learning algorithm to correct the degree of error generated by the entire system.

2.6 Support Vector Regression (SVR)

The regression function $y = f(s)$ for a given dataset $D = \{(s_i, y_i)\}_{i=1}^{n}$, is represented from the perspective of Support Vector Regression (SVR) as a linear function of the form (Wei, Tao, ZhuoShu, & Zio, 2013):

$$f(s) = w^T s + b$$

where w and b are respectively the weight vector and the intercept of the model, and it require to find an optimal fit of the data available in D.

For nonlinear cases, one proceeds by mapping the input low-dimensional vectors via a nonlinear function $\phi = R^p \rightarrow F$, where F is the feature space of ϕ. After nonlinear mapping, the regression function evolves to a pervasive form:

$$f(s) = w^T \phi(s) + b$$

SVR uses the ϵ-insensitive loss function:

$$l = |y - f(s)|_\epsilon = \begin{cases} 0, & |y - f(s)| \leq \epsilon \\ |y - f(s)| - \epsilon, & \text{else} \end{cases}$$

which ignores the error if the difference between the prediction value and the actual value is smaller than ϵ.

By the introduction of ϵ-insensitive loss function, the coefficient w and b can be found by solving a convex optimization problem, which balances the empirical error and the generalization ability. In SVR, the empirical error is measured by means the loss function ϵ-insensitive and the generalization ability is measured by the Euclidean norm of w. Then, the optimization problem to identify the regression model can be formulated by (Wei, Tao, ZhuoShu, & Zio, 2013):

$$\textit{minimize} \quad J(w, \xi_i, \xi_i^*) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} (\xi_i, \xi_i^*)$$

$$\begin{cases} y_i - w^T \phi(s) - b \leq \epsilon + \xi_i \\ w^T \phi(s) + b - y_i \leq \epsilon + \xi_i^* \quad i = 1, \ldots, n \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

where C denotes the penalty parameter between empirical and generalization errors, and ξ_i, ξ_i^* are slack variables how it is observed en the following Figure.

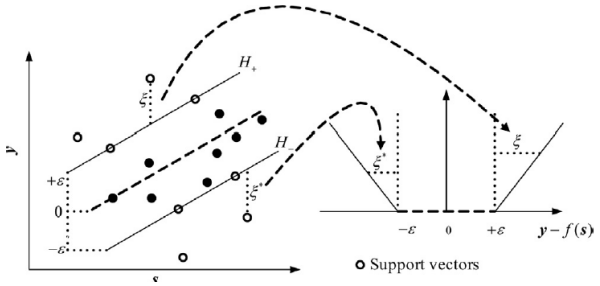

Figure 2: ϵ-insensitive loss function (Wei, Tao, ZhuoShu, & Zio, 2013)

The solution of this optimization problem by the Lagrange method is:

$$f(s) = w^T \phi(s) + b = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(s, s_i) + b$$

where α_i, α_i^* are the Lagrange multipliers of the optimization problem's dual form and $K(s_i, s_j)$ is the kernel function satisfying Mercer condition, and can be described by:

$$K(s_i, s_j) = \phi(s_i)\, \phi(s_j)$$

There are types of kernel functions, the most commons are: linear, polynomial and sigmoid.

3 MATERIALS AND METHODS

3.1 Datasets

The Corporación Bananera Nacional (Corbana) has some research areas, for this research are used data from the 28 Millas research area, it is located at Matina, Costa Rica. The type of plant in study is Banana (*Musa* AAA subgroup Cavendish, cv. Grande Naine).

Although Corbana has meteorological stations that take data every five minutes, for these initial tests it used the weekly averages that generates the same station. The meteorological variables used were:

Table 1: Input variables

Variable	Meaning	Unit
Año	year	2011 thru 2013
Semana	week	1 thru 52
TAireMax	Maximum air temperature	degrees Celsius
TAireMin	Minumum air temperature	degrees Celsius
TempeAire	Average air temperature	degrees Celsius
Humedad	Humidity	percentage
MinHumedad	Minimal humidity	percentage
Rasolwm2	Solar radiation	Watts per square meters
Precipitacion	Average precipitation	millimeters
VelViMax	Maximum wind speed	meters per second
VelViento	Average wind speed	meters per second

With regard to the advancement of the disease, the data were taken weekly and the variable considered is EstadoEvolucion, which is a measure of the level of disease progression (Marin Vargas & Romero Calderón, 1995).

For these first tests and in order to make the regression, although the weather station located at 28 Millas records data from 2003 to the present, they considered the following periods that coincide with the measurement of disease progression: week 01-2006 thru week 46-2013.

3.2 Quality of the Prediction

Although there are many types of indicators to assess the quality of the prediction, the following are used in this research:

Given *n* records, Let be *y* the actual Value of the series and $\hat{y}$ the predicted value.

- MSE: Mean Square Error. $\sum_{i=1}^{n} \frac{(y-\hat{y})^2}{n}$
- RMSE: Root Mean Square Error. $\sqrt{\sum_{i=1}^{n} \frac{(y-\hat{y})^2}{n}}$

The decision to use such indicators is supported by the finding widespread use in the area of machine learning (Soares , Pasqual, & Lacerda, 2013), (Soares, Pasqual, & Lacerda, 2014), (Ibrahim & Wibowo, 2014) and (Demir & Bruzzone, 2014).

The determination of the training set and the test set, especially when the number of records is not very big, authors like Witten (2011) recommend training the model with the available data and using subsets of the data to validate the prediction, this process is called cross validation. It is recommend dividing the total data into 10 parts and run ten times, on each of these parts is used as test set and the prediction error is calculated. Finished the 10 runs, the error average is obtained, which is used like indicator for this model. This process is called tenfold cross-validation validation (Witten, Eibe, & Hall, 2011).

4 RESULTS

4.1 Results Using Sklearn Library

To generate initial results was used python sklearn library (Pedregosa, y otros, 2011) and the numpy library (numpy.org, 2013).

It was tested with the following models: lasso linear model, elasticnet linear model, SVR model with RBF kernel, linear ridge model, bayesian ridge regression model, linear regression model, SVR model with linear kernel.

A total of 396 records were used, 10 folds, training sets of 357 records, cross validation sets of 39 records. The RMSEs obtained were: lasso linear model 502.2553, elasticnet linear model 496.4281, SVR model with RBF kernel 572.6499, linear ridge model 502.2681, bayesian ridge regression model 501.4416, linear regression model 502.2524 and SVR model with linear kernel 790.5167.

4.2 Results Using Echo State Networks

The python-based code used belongs to Dr. Mantas Lukoševičius (2012) from which we made the necessary adjustments for the experiments of this research.

The program was run with different configurations in order to determine the best configuration of neural network. Result: number of neurons: minimum 5, maximum: 145, increases 5; leaking rate: minimum 0.05, maximum 0.95, increments: 0.05; training set: minimum 30, maximum 390, increment: 10; total of records: 394; test set: total of records - training set.

The best configuration was, number of neurons: 45, leaking rate 0.05, training set: 270, test set: 124, MSE: 183470,977 and RMSE: 428.3351.

5 CONCLUSIONS

With the results and considering the quality criteria defined in this document, the models that best learn from data available are: 1) RMSE, Echo State Network: 428.3351, 2) RMSE, Elasticnet Linear model: 496.4281, and 3) RMSE, Bayesian Ridge Regression model: 501.4416.

6 FUTURE WORK

We need to study how to use the greater cardinality of the meteorological data.

It is also among the future work, how to exploit available data through applying other models to work with time series. Between these options are probabilistic graphical models such as Dynamic Bayesian Networks and variants to work on dynamical systems.

With this adjusted models, we can go one-step ahead, to combine data distributed in time and space for forecasting purposes. For example, Corbana has data from another farm.

The ultimate goal of this research is to build a system enabling to learn from data distributed over time and space, and from them to make predictions for the purpose of decision-making in the agricultural field.

7 ACKNOWLEDGEMENTS

The authors would like to thank Corbana for providing the data for this research. Thanks to Dr. Pablo Alvarado Moya, who is the thesis director of Luis Alexánder Calvo.

8 REFERENCES

Camargo, A., Molina, J., Cadena-Torres, J., Jiménez, N., & Kim, J. (2012). Intelligent systems for the assessment of crop disorders. *Computers and Electronics in Agriculture*(85), 1-7. doi:10.1016/j.compag.2012.02.017

Demir, B., & Bruzzone, L. (2014). A multiple criteria active learning method for support vector regression. *Pattern Recognition*, 2558–2567. doi:10.1016/j.patcog.2014.02.001

Huang, Y., Lan, Y., Thomson, S., Fang, A., Hoffmann, W., & Lacey, R. (2010). Development of

soft computing and applications in agricultural and biological engineering. *Computers and Electronics in Agriculture,*(71(2)), 107–127. doi:10.1016/j.compag

Ibrahim, N., & Wibowo, A. (2014). Time Series Support Vector Regression with Missing Data Treatment Based Variables Selection for Water Level Prediction of Galas River in Kelantan Malaysia. *International Journal of Applied Research in Engineering and Science, 3*, 25-36.

Kim, Y., Yoo, S., Gu, Y., Lim, J., Han, D., & Baik, S. (2014). Crop Pests Prediction Method Using Regression and Machine Learning Technology: Survey. *IERI Procedia*(6), 52–56. doi:10.1016/j.ieri.2014.03.009

Lukoševičius, M. (2012). *A minimalistic Echo State Networks demo with Mackey-Glass (delay 17) data*. Retrieved from http://minds.jacobs-university.de/mantas

Lukosevicius, M. (2012). A Practical Guide to Applying Echo State Networks. *Neural Networks: Tricks of the Trade, 7700*. Retrieved from http://reservoir-computing.org/biblio/author/241?sort=type&order=asc

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*(3), 127–149. doi:10.1016/j.cosrev.2009.03.005

Marin Vargas, D., & Romero Calderón, R. (1995). El combate de la Sigatoka Negra. *Boletín Departamento de Investigaciones, Corbana Costa Rica*.

Martín del Brío, B., & Sanz Molina, A. (2010). *Redes neuronales artificiales y sistemas borrosos.* México: Alfaomega Grupo Editor S.A. de C.V.

Murphy, K. P. (2012). *Machine Larning. A probabilistic perspective.* Massachusetts: MIT Press.

numpy.org. (2013). *Sitio web oficial de numpy*. Retrieved from http://www.numpy.org/license.html

Organización de las Naciones Unidas para la alimentación y la agricultura. (2012). *El estado mundial de la agricultura y la alimentación Obtenido desde http://www.fao.org/docrep/017/i3028s/i3028s.pdf.* Retrieved Agosto 19, 2013, from http://www.fao.org/docrep/017/i3028s/i3028s.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Prettenhofer, P. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research, 12*, 2825--2830.

Soares , J., Pasqual, M., & Lacerda, W. (2013). Utilization of artificial neural networks in the prediction of the bunches' weight in banana plants. *Scientia Horticulturae*(155), 24-29.

Soares, J., Pasqual, M., & Lacerda, W. (2014). Comparison of techniques used in the prediction of yield in banana plants. Scientia Horticulturae, 167, 84–90. doi:10.1016/j.scienta.2013.12.012. *Scientia Horticulturae journal, 167*, 84-90.

Wei, Z., Tao, T., ZhuoShu, D., & Zio, E. (2013). A dynamic particle filter-support vector regression method for reliability prediction. *Reliability Engineering & System Safety*, 109–116. doi:10.1016/j.ress.2013.05.021

Witten, I., Eibe, F., & Hall, M. (2011). *Data Mining. Practical machine learning tools and techniques*. USA: Morgan Kaufmann Publiisher.

Zhao, L., He, L., Harry, W., & Jin, X. (2013). Intelligent Agricultural Forecasting System Based on Wireless Sensor. *Journal of Networks*(8), 1817–1824. doi:10.4304/jnw.8.8.1817-1824