Early warning system for coffee rust disease based on error correcting output codes: a proposal

David Camilo Corrales[*]
Andrés J. Peña Q.[**]
Carlos León[***]
Apolinar Figueroa[****]
Juan Carlos Corrales[*****]

Recibido: 25/10/2013 • Aceptado: 27/06/2014

Abstract

Colombian coffee producers have had to face the severe consequences of the coffee rust disease since it was first reported in the country in 1983. Recently, machine learning researchers have tried to predict infection through classifiers such as decision trees, regression Support Vector Machines (SVM), non-deterministic classifiers and Bayesian Networks, but it has been theoretically and empirically demonstrated that combining multiple classifiers can substantially improve the classification performance of the constituent members. An Early Warning System (EWS) for coffee rust disease was therefore proposed based on Error Correcting Output Codes (ECOC) and SVM to compute the binary functions of Plant Density, Shadow Level, Soil Acidity, Last Nighttime Rainfall Intensity and Last Days Relative Humidity.

Key words: Coffee Rust Disease, Early Warning System, ECOC, SVM, Codeword.

[*] M.Sc. in Telematics Engineering, and Researcher of Telematics Engineering Group and Environmental Study Group at University of Cauca, Colombia. Address. Carrera 2 # 1A-25. Popayán, Colombia. Tel. +57 (8) 209800 Ext. 2145. E-mail. dcorrales@unicauca.edu.co

[**] M.Sc. in Meteorology, and Researcher at National Coffee Research Center (Cenicafé), Colombia. Address: Chinchiná Km 4 via. Manizales, Colombia. Tel. +57 (6) 8506550. E-mail. andres.pena@cafedecolombia.com

[***] M.Sc. in Electrical Engineering, and CEO of CreaTIC Corporation - Parquesoft, Colombia. Address: Calle 17 N # 6 – 53, Popayán, Colombia. Tel. +57 3017917388. E-mail. cleon65@corporacioncreatic.com

[****] Doctor of Biological Sciences, and Full Professor and Leader of the Environmental Study Group at University of Cauca, Colombia. Address: Carrera 2 # 1A-25. Popayán, Colombia. Tel. +57 (8) 209800 Ext. 2145. E-mail. apolinar@unicauca.edu.co

[*****] Doctor of Philosophy in Sciences, Speciality Computer Science, and Full Professor and Leader of the Telematics Engineering Group at University of Cauca, Colombia. Address: Calle 5 No. 4 - 70. Popayán, Colombia. Tel. +57 (8) 209800 Ext. 2129. E-mail. jcorral@unicauca.edu.co

Sistema de alerta temprana para la roya en el café basado en códigos de salida de corrección de error: una propuesta

Resumen

Los productores de café colombianos han sufrido severas consecuencias por la Roya desde que fue reportada por primera vez en el país en el año 1983. Recientemente, investigadores de aprendizaje automático han intentado predecir la roya a través de clasificadores como: arboles de decisión, máquinas de vector de soporte, clasificadores no determinísticos y redes bayesianas, pero se ha demostrado teórica y empíricamente que la combinación de múltiples clasificadores puede mejorar sustancialmente el rendimiento en la clasificación. En este sentido es propuesto un sistema de alerta temprana para la roya en el café, basado en códigos de salida de corrección de error y máquinas de vector de soporte para calcular las funciones binarias de la densidad de planta, el nivel de sombra, la acidez del suelo, la intensidad de lluvia en la última noche, y en últimos días, con humedad relativa.

Palabras clave: roya, sistema de alerta temprana, ECOC, SVM, Codeword.

INTRODUCTION

Coffee production is the main agricultural activity in Colombia. More than 350.000 families depend on coffee harvest for their sole income. As such, diseases, pests and even low prices impact terribly on the economic and social aspects of the main coffee-growing regions. Coffee rust, first reported in 1983 [1], is the most important and severe disease currently affecting the production of Colombian coffee. Resistant varieties have been developed through improvement with genes of Timor Hybrid (a plant that features natural resistance to the disease) as a solution to the rust problem [2], yet more than 50 percent of the country's coffee crop is still susceptible in the productive phase. Studies on coffee rust have concluded that the spores carrying the infection are spread by climatic elements such as wind and rainfall [3], wind being the vector for long distance spore transport, while precipitation droplets are responsible for vertical propagation from infected leaves or soil [3]. Once spores make contact with a susceptible leaf, the infection process is improved by high shadow index, high humidity (atmosphere and leaf), soil acidity, high coffee tree density and low soil fertility. The warning system proposed herein seeks to detect each of the favorable conditions that coffee rust requires to infect the crop and, by taking prophylactic measures (biological, chemical and cultural control), thus allow prevention of the onset of the disease. Having identified the particular conditions favorable to coffee rust, the system is designed to alert growers to the settled presence of conditions for infection, sending out a graded warning (None, Very Low, Low, Medium, High, Very High) to all coffee growers in the detection area.

Machine learning researchers have in recent times attempted to predict infections through such classifiers as: decision trees [4], regression Support Vector Machines [5], non-deterministic classifiers [6] and Bayesian Networks [7]. However, it has been demonstrated theoretically and empirically that using combinations of multiple classifiers can substantially improve upon the performance of constituent members [8]. It was therefore proposed using an ECOC to alert about coffee rust disease. This method is considered excellent for applying binary learning algorithms to multiclass learning problems [9]. The rest of the paper is organized as follows: the SVM classifier and ECOC method are introduced in Section and Section respectively. The final proposal of ECOC in alerting to the risk of coffee rust disease is presented in Section 61, while Section 63 reports a number of conclusions.

1. SUPPORT VECTOR MACHINES FOR BINARY CLASSIFICATION

Support vector machines (SVM) constitute an effective binary data classification method. The key idea of SVMs is the use of a mapping function that projects the given input feature space into a high dimensional feature space to find an optimal hyperplane,

having the largest margin of separation between different classes with minimum error rate. SVMs use a portion of the data to train and find several support vectors that represent the training data. These support vectors are then formed into a model by the SVM, representing each category. For a linearly separable binary classification with an n-dimensional vector x_i and the label of the class that vector y_i, i.e., $\{(x_i, y_i)\}_{i=1}^{N}$ and $y_i = \{+1, -1\}$, the SVM separates the two classes of points using the classification decision function $f_{w,b} = sign(w \cdot x + b)$, where w is an input vector, x is an adaptive weight vector, and b is a bias. SVM finds the parameters w and b for the optimal hyperplane to maximize the geometric margin [10],

$$\frac{2}{w}, \text{ subject to min } \left(\frac{w^T w}{2}\right), y_i(w \cdot x + b) \geq +1 \tag{1}$$

2. ERROR CORRECTING OUTPUT CODES (ECOC)

ECOC is a powerful approach to dealing with multi-class problems based on the combination of binary classifiers. The ECOC approach works in two steps [10]:

1. **The coding step.** In this step, a set of B different bipartitions of the class label set $\{C_1, \ldots, C_l\}$ are constructed, and subsequently B binary classifiers $h_1, \ldots, h_B$ are trained over the partitions.
2. **The decoding step.** In this step, given an instance x, a **codeword** is generated by using the outputs of B binary classifiers. The codeword is then compared to the base codeword of each class, and the instance is assigned to the class with the most similar codeword. The central task of decoding is to find the base codeword w_i (corresponding to class c_i) which is the closest to the codeword v of the given test instance. The binary decoding scheme most used is presented next:

Hamming decoder (HD): This scheme is based on the assumption that the learning task can be modeled as an error-correcting communication problem [11]. The measure is given by:

$$HD(v, w_i) = \frac{\sum_j \left(1 - \text{sign}\left(v^j \cdot w_i^j\right)\right)}{2} \tag{2}$$

Typically, the partitions of the class set are specified by a coding matrix *M*, which can appear in two forms, binary form and ternary form. In this paper use of the binary form is proposed, where $M \in \{-1, +1\}^{lxB}$. Fig. 1 provides an example of a binary coding matrix, which transforms a four-class problem into five binary classification problems. In the figure, the regions coded by +1 are considered as a class, while regions coded by -1 are considered as another class [10]. Consequently, the binary classifiers

selected for this paper are Support Vector Machine, considered the most accurate for classification problems [12]. This classifier is presented in section 2.

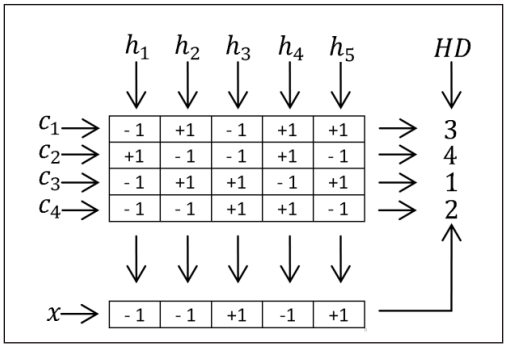
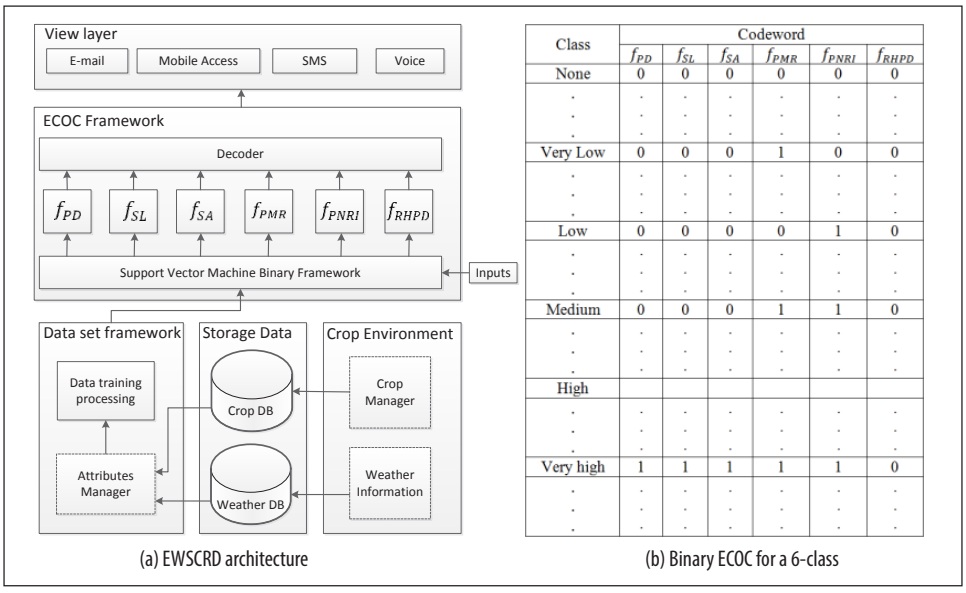


Fig. 1. Binary ECOC for a 4-class problem. An instance $\boldsymbol{x}$ is classified to class $\boldsymbol{c}_3$ using the Hamming (HD) (Source:[10])

3. EARLY WARNING SYSTEM FOR COFFEE RUST DISEASE (EWSCRD)

The EWSCRD approach detects all favorable conditions that coffee rust requires to infect a crop. These favorable conditions are identified through input data from coffee growers or from agro-meteorological station data. The architecture proposed for EWSCRD consists of five components: Crop environment, Storage data, Data set framework, ECOC framework, and View layer. These are shown in Fig. 2 (a), and are explained in the following:



Class	Codeword					
	f_{PD}	f_{SL}	f_{SA}	f_{PMR}	f_{PNRI}	f_{RHPD}
None	0	0	0	0	0	0
Very Low	0	0	0	1	0	0
Low	0	0	0	0	1	0
Medium	0	0	0	1	1	0
High						
Very high	1	1	1	1	1	0

(b) Binary ECOC for a 6-class

Fig. 2. (a) Architecture; (b) Binary ECOC for a 6-class, includes 64 examples.

– **Crop enviroment (CE).** Consists of two sub-components: the weather information sub-component, which accesses weather data (temperature, rainfall, etc.) from the agro-meteorological station network and farmer observation, and the crop manager subcomponent, which compiles detailed crop information (sowing spacing between plants, plant shadows, soil pH) through the data inserted by farmer.

– **Storage data (SD).** Once information is captured in the weather and crop subcomponets of CE, the data are stored in the weather database (Weather DB) and crop database (Crop DB) respectively.

– **Data set framework (DSF).** This component creates the training data from two subcomponents: the attributes manager, where an expert user defines the attributes and classes for building the data training, and the data training processing subcomponent which transforms the data from the Weather DB, Crop DB and the attributes manager subcomponent to the data training set.

– **ECOC framework (ECOCF):** The ECOCF receives the data training built in DFS in order to train the SVM binary framework (SVMBF) subcomponent. In this sense, each function $f_{PD}, f_{SL}, f_{SA}, f_{PMR}, f_{PNRI}, f_{RHPD}$ is considered a binary classifier (Fig. 2 (b)) built through SVMBF, which solves the following problems:

  - **Plant density (f_{PD}) :** related to the number of coffee trees per unit area. Less than 6,500 trees per hectare is considered low density (0). More than 6,500 trees is considered high (1).

  - **Shadow level (f_{SL}):** related to the presence of trees (larger than the coffee crop) in the unit area. If the shade provided by these is greater than 25% it is considered a high level of shade (1). Below 25% is considered a low level of shade (0).

  - **Soil acidity (f_{SA}):** related to low soil pH. Soil pH values below 5.5 are considered acidic (1). Values greater than 5.5 are considered non-acidic (0).

  - **Previous month rainfall (f_{PMR}):** related to leaf moisture in the previous month. When it has rained more than 120 millimeters in the previous month, the rainfall is considered high (1); less than this threshold is considered low (0).

  - **Previous nighttime rainfall intensity (f_{PNRI}):** related to the extent that the leaf and the canopy of the plantation was wet during the previous week. If there were more than three days in the week with rain at night, it is considered high (1), while if the number of days with rain at night was less than three, it is considered low (0).

- **Relative humidity in the previous days (f_{RHPD}):** related to the atmosphere and its ability to dry the leaves. Ten days with relative humidity over 85% promotes the disease (1); less than 85% does not promote the disease (0).

The combination the outputs for each function (0 or 1) generates a codeword x (Fig. 2 (b)) in the decoder subcomponent. The central task of this subcomponent is to find the codeword corresponding to one of the classes (None, Very Low, Low, Medium, High, Very High) that is the closest to the codeword x. Thus, the process for choosing the most similar codeword is based on computing the Hamming distance.

– **View Layer:** shows the results obtained in ECOCF through various devices such as mobile phone (SMS), e-mail, voice, etc.

4. CONCLUSIONS

While coffee rust is the most important coffee crop disease in Colombia, local growers have no access to early warning systems (EWS) for management of the disease. Studies have been done relating to the design of just such systems, but in most cases proposals have been based on linear approximations (univariate) that connect some environmental factor (e.g. temperature, relative humidity) to the disease [13]. As a consequence, the models are not very accurate and when the farmer comes to use the system it is possible that he does not have the input data relating to the relevant environmental factor. Ideally, the farmer should need only to have access to a system of alert. With this in mind, ECOC is considered excellent for applying binary learning algorithms to multiclass learning problems [9] and SVM is the most accurate algorithm for classification problems [12]. As such, ECOC + SVM may be a powerful tool to solve the particular problem of coffee rust infection. The proposed EWSCRD gives the farmer this alert option. More than this, EWSCRD is easy to use in any coffee growing scenario, whether meteorological data is available or not – this latter being the reality in Colombia for the most part. In other words, with EWSCRD the absence of models relating climate, management, and soil with production, pests and diseases e.g. for coffee rust in Colombia, can be solved in that the alert is related to the presence (or absence) of identified events (climate, soil or management). The importance is seen in avoidance of coffee rust disease and a consequent vital enhancement in coffee growing income.

5. ACKNOWLEDGEMENTS

This paper was presented in the 1st CSIRO Workshop on 'Semantic Machine Learning and Linked Open Data (SML2OD2013) for Agricultural and Environmental Informatics' was held in conjunction with the 12th International Semantic Web Conference

(ISWC 2013), on 22nd October 2013, in Sydney, Australia. The authors are grateful to the Environmental Study Group (GEA), the Telematics Engineering Group (GIT) of the University of Cauca, and AgroCloud project of the RICCLISA Program for technical and scientific support.

REFERENCES

[1] E. Shieber and G. A. Zentmyer, "Coffee rust in the western hemisphere " in *Plant disease*, ed, 1984.

[2] J. C. Zapata, G. M. Ruíz, F. N. d. C. d. Colombia, and C. N. d. I. d. Café, *La variedad Colombia: selección de un cultivar compuesto resistente a la roya del cafeto : Premio Nacional de Ciencias Fundacón Alejandro Angel Escobar, 1986*: Cenicafé, 1988.

[3] S. Becker, in *La propagación de la roya del cafeto*, ed: Sociedad alemana de cooperación técnica ltda. (GTZ), 1979, p. 70.

[4] M. E. Cintra, C. A. A. Meira, M. C. Monard, H. A. Camargo, and L. H. A. Rodrigues, "The use of fuzzy decision trees for coffee rust warning in Brazilian crops," in *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, 2011, pp. 1347-1352.

[5] O. Luaces, L. H. A. Rodrigues, C. A. A. Meira, Jos, #233, R. Quevedo*, et al.*, "Viability of an alarm predictor for coffee rust disease using interval regression," presented at the Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part II, Cordoba, Spain, 2010.

[6] O. Luaces, L. H. A. Rodrigues, C. A. Alves Meira, and A. Bahamonde, "Using nondeterministic learners to alert on coffee rust disease," *Expert Systems with Applications,* vol. 38, pp. 14276-14283, 2011.

[7] C. B. Pérez-Ariza, A. E. Nicholson, and M. J. Flores, "Prediction of Coffee Rust Disease Using Bayesian Networks," presented at the The Sixth European Workshop on Probabilistic Graphical Models, Granada (Spain), 2012.

[8] L. Li, B. Zou, Q. Hu, X. Wu, and D. Yu, "Dynamic classifier ensemble using classification confidence," *Neurocomputing,* vol. 99, pp. 581-591, 2013.

[9] T. G. Dietterich and G. Bakiri, "Error-correcting output codes: a general method for improving multiclass inductive learning programs," presented at the Proceedings of the ninth National conference on Artificial intelligence - Volume 2, Anaheim, California, 1991.

[10] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*: Chapman and Hall/CRC;, 2012.

[11] N. J. Nilsson, *Learning machines: foundations of trainable pattern-classifying systems*. New York: McGraw-Hill, 1965.

[12] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," presented at the Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007.

[13] G. C. Gómez, "Diseminación de Hemileia vastatrix Berk y Br.," in *Taller sobre Roya del cafeto*, ed Manizales: Federación Nacional de Cafeteros - Centro Nacional de Investigaciones de Café, 1982, pp. 1 - 27.