Estimation of coffee rust infection and growth through two-level classifier ensembles based on expert knowledge

David Camilo Corrales*

Grupo de Ingeniería Telemática,
Universidad del Cauca,
Campus Tulcán, Popayán, Colombia
and
Departamento de Ciencias de la Computación e Ingeniería,
Universidad Carlos III de Madrid,
Avenida de la Universidad 30, 28911 – Leganés, Spain
Email: dcorrales@unicauca.edu.co
Email: davidcamilo.corrales@alumnos.uc3m.es
*Corresponding author

Emmanuel Lasso

Grupo de Ingeniería Telemática,
Universidad del Cauca,
Campus Tulcán, Popayán, Colombia
Email: eglasso@unicauca.edu.co

Apolinar Figueroa Casas

Grupo de Estudios Ambientales,
Universidad del Cauca,
Carrera 2 No. 1A-25 – Urbanización Caldas, Popayán, Colombia
Email: apolinar@unicauca.edu.co

Agapito Ledezma

Departamento de Ciencias de la Computación e Ingeniería,
Universidad Carlos III de Madrid,
Avenida de la Universidad 30, 28911 – Leganés, Spain
Email: ledezma@inf.uc3m.es

Juan Carlos Corrales

Grupo de Ingeniería Telemática,
Universidad del Cauca,
Campus Tulcán, Popayán, Colombia
Email: jcorral@unicauca.edu.co

Abstract: Rust is a disease that leads to considerable losses in the worldwide coffee industry. There are many contributing factors to the onset of coffee rust, e.g., crop management decisions and the prevailing weather. In Colombia the coffee production has been considerably reduced by 31% on average during the epidemic years compared with 2007. Recent research efforts focus on detection of disease incidence using computer science techniques such as supervised learning algorithms. However, a number of different authors demonstrate that results are not sufficiently accurate using a single classifier. Authors in the computer field propose alternatives for this problem, making use of techniques that combine classifier results. Nevertheless, the traditional approaches have a limited performance due to dataset absence. Therefore, we proposed two-level classifier ensembles for coffee rust infection and growth estimation in Colombian crops, based on expert knowledge.

Keywords: coffee; rust; classifier; ensemble; dataset; expert; knowledge.

Reference to this paper should be made as follows: Corrales, D.C., Lasso, E., Casas, A.F., Ledezma, A. and Corrales, J.C. (2018) ‘Estimation of coffee rust infection and growth through two-level classifier ensembles based on expert knowledge’, *Int. J. Business Intelligence and Data Mining*, Vol. 13, No. 4, pp.369–387.

Biographical notes: David Camilo Corrales received his degree in Informatics Engineering and Master in Telematics Engineering from the University of Cauca, Colombia, in 2011 and 2014 respectively. He is a PhD scholarship holder of Colciencias in Telematics Engineering at the University of Cauca and Science and Informatics Technologies at Carlos III of Madrid University. His researches focus on artificial intelligence, data mining, machine learning and data analysis.

Emmanuel Lasso received his degree in Electronics and Telecommunication Engineering and Master in Telematics Engineering from University of Cauca, Colombia, in 2013 and 2016 respectively. He is a PhD scholarship holder of Innovaccion Cauca in Telematics Engineering at the same institution. His researches focus on graphs, knowledge representation and management, data mining, machine learning and decision making.

Apolinar Figueroa received his degree in Biology from University of Cauca, Colombia, in 1982, Master’s degree in Ecology from University of Barcelona, Spain, in 1986, and PhD in Biological Sciences from University of Valencia, Spain, in 1999. Currently, he is a Full Professor and leads the Environmental Studies Group at University of Cauca. His research interests focus on environmental impact assessment and biodiversity management.

Agapito Ledezma is an Associate Professor in the Department of Computer Science at Carlos III of Madrid University. He received his BS from Universidad Latinoamericana de Ciencia y Tecnologa in 1997 (Panama, Panama) and his PhD in Computer Science from Carlos III University in 2004 (Madrid, Spain). His research interests centre on machine learning, activity recognition, intelligent agents and advanced driving assistant systems. He has published over 80 journal and conference papers mainly in the field of artificial intelligence and machine learning.

Juan Carlos Corrales received his Dipl-Ing and Master’s degree in Telematics Engineering from the University of Cauca, Colombia, in 1999 and 2004 respectively, and PhD in Sciences, specialty Computer Science, from the

University of Versailles Saint-Quentin-en-Yve-lines, France, in 2008. Currently, he is a Full Professor and leads the Telematics Engineering Group at the University of Cauca. His research interests focus on service composition and data analysis.

This paper is a revised and expanded version of a paper entitled 'An empirical multi-classifier for coffee rust detection in Colombian crops', presented at International Conference on Computational Science and Its Applications, ICCSA 2015: Computational Science and Its Applications – ICCSA 2015, Banff, AB, Canada, 19 June 2015.

1 Introduction

Coffee rust is a leaf disease caused by the fungus, Hemileia vastatrix. Coffee rust epidemics, with intensities higher than previously observed, have affected a number of countries including: Colombia, from 2008 to 2011; Central America and Mexico, in 2012–2013; and Peru and Ecuador in 2013. There are many contributing factors to the onset of these epidemics, e.g., the state of the economy, crop management decisions and the prevailing weather, and many resulting impacts, e.g., on production, on farmers and labourers income and livelihood, and on food security. Production has been considerably reduced in Colombia (by 31% on average during the epidemic years compared with 2007) and Central America (by 16% in 2013 compared with 2011–12 and by 10% in 2013–2014 compared with 2012–2013) (Avelino et al., 2015).

Since coffee rust has led to considerable losses in the industry worldwide, recent Brazilian supervised learning researchers have focused on estimation of the incidence of the disease using simple classifiers as decision trees, support vector machines (SVMs) and Bayesian networks (Cintra et al., 2011; Luaces et al., 2010, 2011; Meira et al., 2008, 2009; Pérez-Ariza et al., 2012). They relate the different variables involved in the development of coffee rust, identified based on expert knowledge, in order to generate predictive models.

Meanwhile, computer science experts demonstrated that using a simple classifier is not accurate enough (Li et al., 2013). In this sense, several authors suggested an alternative solution to make use of techniques that combine classifier results (Ghosh, 2002; Ranawana and Palade, 2006) named ensemble methods. Indeed, the authors of this research have proposed the use of a multi-classifier for detecting coffee rust incidence (Corrales et al., 2015a). However, it is not considered predicting the disease growth rate in order to characterise a tendency of it. In the same way, it is not taken into account the expert knowledge in coffee rust to build attributes within the training set used to generate the classifiers.

Therefore, we proposed two-level classifier ensembles for coffee rust infection and growth estimation in Colombian crops, based on expert knowledge. The remainder of this paper is organised as follows: Section 2 describes the data collection and the algorithms used; Section 3 the algorithms used in the two-level classifier ensembles proposed; Section 4 presents results and discussion and Section 5 conclusions and future work.

2 Background

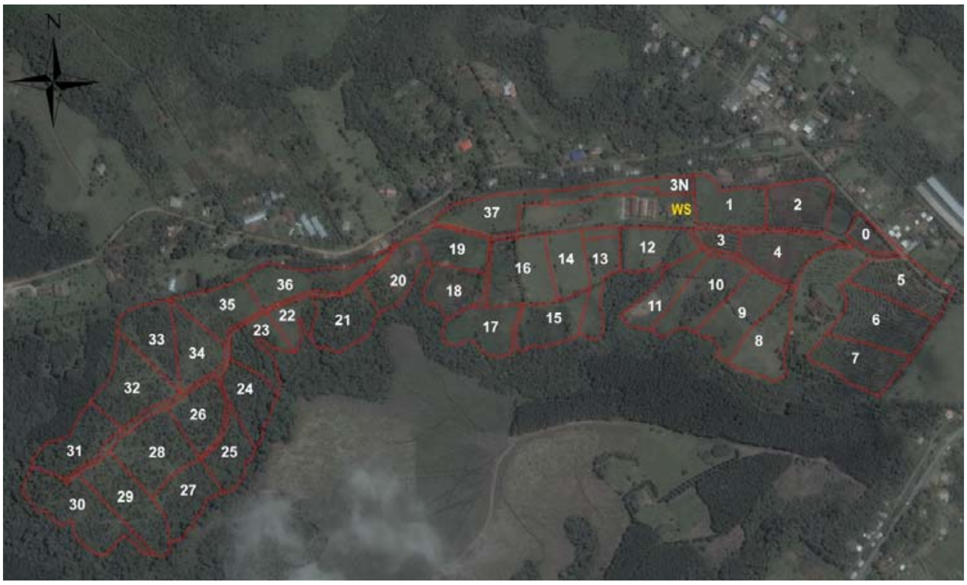
This section describes the expert knowledge in coffee rust, data collection process and the generation of the dataset used in experiments.

2.1 Expert knowledge in coffee rust

Coffee rust has been studied by several researchers around the world, considering its large negative effects on coffee crops. Thus, they have found an association between outbreaks of this disease and factors such as: physic crop properties, crop management and the distribution of some weather conditions (rainfall, temperature and relative humidity) (Rivillas et al., 2011; Waller et al., 2007).

There are considerations for each variable that affects the life cycle of the fungus, which can draw from the expert knowledge (Lasso et al., 2015). Crop density determines the competition between plants for nutrients, spore interception and coverage of fungicides on the foliage (Rivillas et al., 2011); while excessive shade increases the infection intensity (Nutman et al., 1963). At the same time, the fungus requires splatter rain to begin the process of dispersion, as well as the presence of a layer of water on the underside of leaves to germinate (Rivillas et al., 2011). To estimate the period of leaf wetness, is measured the number of hours with relative air humidity above a specific limit, usually 90% or 95%. (Sutton et al., 1984). Moreover, six hours of leaf wetness was established as the minimum time required for an infection occurs (Kushalappa et al., 1983). In the same way, once the leaf surface is wet, the temperature is the main factor that determines the percentage of spore germination and penetration (Kushalappa et al., 1983). Temperatures between 16 and 28°C directly influence the development of coffee rust (Rivillas et al., 2011). Finally, wind is a required element of the fungus dispersion (Becker, 1977).

Figure 1 Technical farm: Los Naranjos (see online version for colours)



2.2 Data collection

The data used in this work were collected (Corrales et al., 2014) trimonthly for 18 plots (1–17 and 37, in the Figure 1), closest to weather station at the technical farm (Naranjos) of the Supracafé, in Cajibio, Cauca, Colombia (21°35’08”N, 76°32’53”W), during the last three years (2011–2013). The dataset includes 147 examples from the total of 216 available ones. The remaining 69 samples were discarded due to problems in the collection process.

The dataset is composed of 13 attributes, taking into account the expert knowledge in coffee rust presented in the previous section, and is divided in three categories: weather conditions (six attributes), physic crop properties (three attributes), and crop management (four attributes). Weather conditions correspond to weather station monitoring data and the remaining information was extracted from crop control and management registers made by farm owners. Below are described the 13 attributes (Table 1).

Table 1 Dataset for incidence rate of rust (IRR) estimation

Attributes for IRR detection	
Weather conditions	Temperature, relative humidity, precipitation, and some special variables as: relative humidity average in the last two months (RHA2M), hours of relative humidity > 90% in the last month (HRH1M), temperature variation average in the last month (TVA1M), hourly precipitation, rainy days in the last month (RD1M), accumulated precipitation in the last two months (AP2M), nightly accumulated precipitation in the last month (NAP1M).
Physic crop properties	Coffee variety (CV), crop age (CA), percentage of shade (PS).
Crop management	Coffee rust control in the last month (CRC1M), coffee rust control in the last three months (CRC3M), fertilisation in the last four months (F4M), accumulated coffee production in the last two months (ACP2M).

In this sense, the class was defined as, the IRR. IRR is calculated by following a unique methodology in Colombian coffee crops collection developed by Cenicafé (Rivillas-Osorio et al., 2011) for a plot with area lower or equal of one hectare. The steps of the methodology are presented below:

1 The farmer must be standing in the middle of the first furrow and he has to choose one coffee tree and pick out the branch with greater foliage for each level (high, medium, low); the leaves of the selected branches are counted as well as the infected ones for rust.

2 The farmer must repeat the step 1 for every tree in the plot until 60 trees are selected. Take in consideration that the same number of trees must be selected in every furrow (e.g., if plot has 30 furrows, the farmer selects two coffee trees for each furrow).

3 Finished the step 1 and 2, the leaves of the coffee trees selected (*LCT*) are added as well as the infected leaves of rust (*ILR*). Later it must be computed the IRR using the following formula:

$$IRR = \frac{ILR}{LCT} \tag{1}$$

IRR can take a value between 0 and 1, so we take it as a percentage of IRR (multiplying it by 100). The collection process and IRR computation spend large amount of money and time, for this reason the IRR samples are limited (made quarterly for 18 plots). This process and its samples are considered very important, since it provides coffee crops rust approximation.

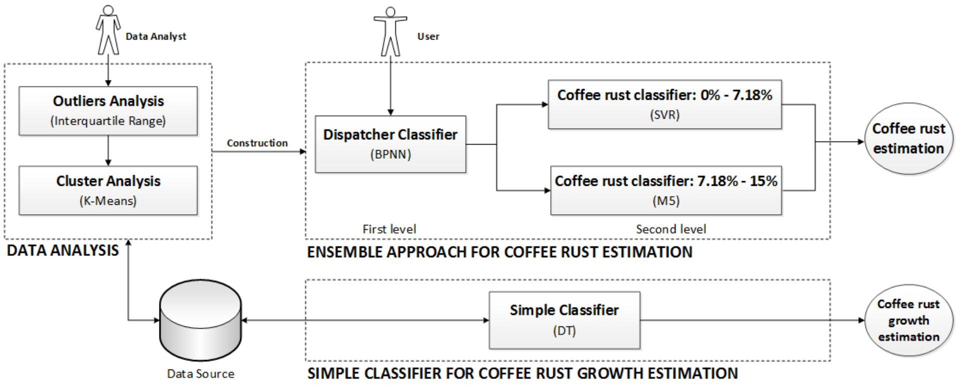
3 Our approach

What we want to do is to recognise the state of a crop against coffee rust, based on the monitoring of climate data and its physical properties. For this purpose, we are using an ensemble approach for coffee rust infection estimation and a simple classifier to estimate the disease growth. Thus, the combination of the two results can be seen as a prediction-information warning, relating the estimate of the actual disease incidence (infection) and its tendency to increase or decrease (growth).

The ensemble approach for coffee rust infection estimation is focused on the use of a simple classifier (decision tree DT) for disease growth estimation and multiple classifiers to generate a disease infection level estimation, so that each of these classifiers (back propagation neural network BPNN, regression tree M5 and support vector regression SVR) covers a different part of the dataset. It is important to highlight that we want to integrate the classification results (from simple and multiple classifiers) and generate the final classification. In addition, we used the interquartile range and k-mean algorithms to improve the performance in the dataset in data analysis module.

In Figure 2 it can be seen the integration scheme of the two approaches and their response characterisation. Data analysis component will be explained in Section 3.1, the ensemble approach for coffee rust estimation in Section 3.2 and the simple classifier for coffee rust growth estimation in Section 3.3.

Figure 2 Workflow of two-level classifier ensembles for coffee rust infection and growth estimation based on expert knowledge



In this way, the final user types the inputs (weather conditions, physic crop properties, and crop management) and the ensemble approach returns the estimation of infection rate of rust, while the simple classifier returns an estimation of coffee rust growth.

3.1 Data analysis

This module examines data which deviate so much from other data (Corrales et al., 2015b) through outliers analysis sub-module, besides data analysis module (DA) defines the amount of base classifiers of the ensemble approach for coffee rust estimation supported by cluster analysis sub-module. This module is handled by an expert in data analysis such as: data scientist, data analyst, data miners, etc.

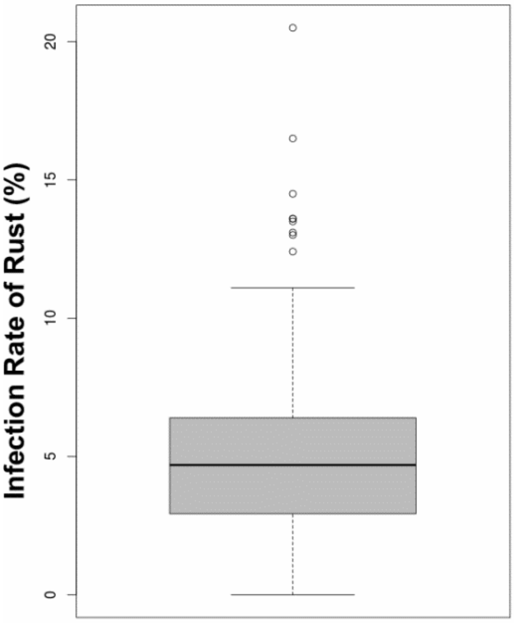
3.1.1 Outliers analysis

The outlier's analysis sub-module detects and removes the values that have an abnormal behaviour of an attribute into the dataset through interquartile range method (IR). This method compute the quartile Q_1, Q_2, Q_3, which split a sort dataset in four parts (McAlister, 1879), after, computing the interquartile range (IQR), which is the difference among the third quartile (Q_3) and first quartile (Q_1). The IQR is a measure of noise for the data set. Points that are beyond the quartiles by half IQR's will be deemed potential outliers (Grubbs, 1969). Below is presented the mathematic representation:

$$x < q_1 - 1.5xIQR \vee x > q_3 + 1.5xIQR; \forall x \in \mathbb{R} \tag{2}$$

where, x is the observation to evaluate and $q_1 - 1.5xIQR$ denotes the lower inner fences and $q_3 + 1.5xIQR$ the upper inner fences. Hence points beyond these fences are potential outliers. In this case the interquartile range method was applied to the class – IRR. The outcomes obtained to apply IR, presented as lower inner fence –2.0156 and upper inner fence 11.84, thus, are removed nine observations with values of IRR: 12.41%, 13.01%, 13.1%, 13.5%, 13.6%,13.6%, 14.5%, 16.5%, 20.50% greater than upper inner fence 11.84% as we can see in Figure 3.

Figure 3 Box-plot of infection rate of rust



3.1.2 Cluster analysis

The cluster analysis sub-module builds clusters from data set (leaving out the IRR class) using the well known and simple clustering algorithm K-means (Jain, 2010). This algorithm partitions a set of data into a number k of disjoint clusters by looking for inherent patterns in the set. Let us suppose that X represents the available set of samples. Each sample can be represented by an m-dimensional vector in the Euclidean space R^m. Thus, in the following, $X = \{x_1, x_2,\ldots, x_3,\}$ will represent a set of n samples, where the generic sample x_i is a m-dimensional vector (Mucherino et al., 2009). Each cluster is a subset of X and contains samples with some similarity. The distance between two samples provides a measure of similarity: it shows how similar or how different two samples are. In the k-means approach, the representative of a cluster is defined as the mean of all the samples contained in the cluster (Mucherino et al., 2009).

3.1.3 Interpretation of clusters

Once the k-means is applied, each cluster is transformed into a data training set. The basic idea is to extract the meaning of the clusters and define an expert classifier for each cluster. Next is explained the interpretation of the clusters through Bayesian network and decision tree.

We used a Bayesian network (Araujo, 2006) to build a conditional probability distribution for clusters generated by k-means (k = 3, as explained below) and four main attributes: coffee rust control in the last month (CRC1M), coffee rust control in the last three months (CRC3M), fertilisation in the last four months (F4M) and IRR, as can be seen in Table 2.

Table 2 Percentage of conditional probability distribution for clusters generated by k-means

Cluster	*CRC1M*		*CRC3M*		*F4M*		*IRR*	
	Yes	*Not*	*Yes*	*Not*	*Yes*	*Not*	*< 7.18%*	*≥ 7.18%*
C1	45.3%	54.7%	98.3%	1.7%	32.8%	67.2%	75%	25%
C2	1%	99%	27.9%	72.1%	20.2%	79.8%	66.66%	33.33%
C3	88.6%	11.4%	58.8%	41.2%	99.1%	0.9%	98.70%	1.30%

For cluster C_1 , when it was done a coffee rust control in the last month and three months (45.3% – CRC1M = 'yes' and 98.3% – CRC3M = 'yes'), exists a probability distribution unequal and incoherent instances, because if the majority of instances of CRC3M shows that it was done a coffee rust control, mandatorily the instances of CRC1M must present the same behavioural.

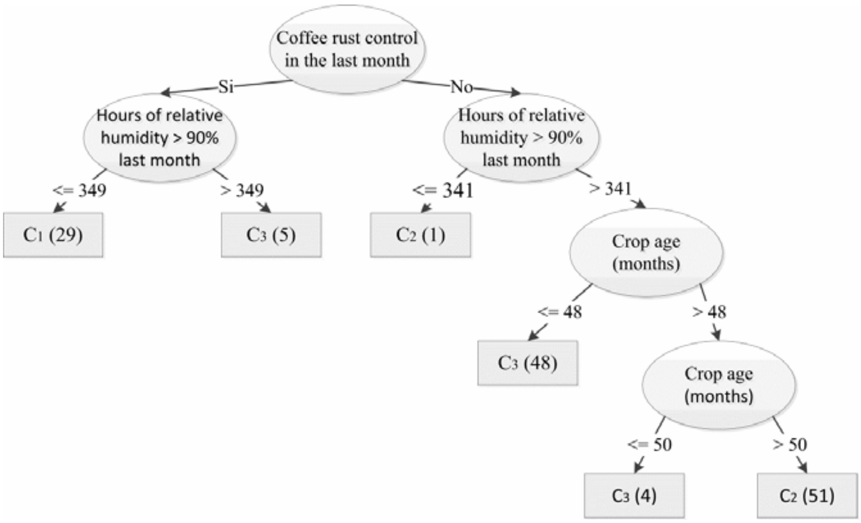
On the other hand, if we analyze the cluster C_2, the attributes of CRC1M, CRC3M and F4M present a similar behaviour, which can be interpreted as the instances in C_2 have not done rust controls and fertilisations on coffee crops (probability distribution of 'not' values: 99%, 72.1% and 79.8% respectively), whereas probability distribution of CRC1M, CRC3M and F4M attributes of C_3 indicates the use of coffee rust controls and fertilisations (probability distribution of 'yes' values: 88.6%, 58.8% and 99.1% respectively); for this reason the IRR is less than 7.18% (probability distribution of IRR < 7.18% = 98.70%).

To test out the outcomes obtained by Bayesian network, we used a C4.5 decision tree (pruning the irrelevant attributes) (Quinlan, 1993), as can be seen in Figure 4.

In Figure 4, we can observe that the C4.5 decision tree accounted three attributes: coffee rust control in the last month (CRC1M), hours of relative humidity > 90% in the last month (HRH1M), and crop age (CA) in months. The distribution of instances is founded in the leaves. In this sense the rule obtained for cluster C_1 (Figure 4) does not contain the necessary attributes to know the meaning of C_1 (CRC1M and HRH1M). However, in C_2 the conditions promote the appearance of rust, because there was not done a coffee rust control in the last month, high hours of relative humidity > 90% in the last month (> 341 hours), and older crops (age > 50 months). Finally C_3 can be interpreted as the youngest crops (age < 48 months) that are resistant to rust without regard for conditions of relative humidity and coffee rust controls performed.

Based on the foregoing, we interpret the cluster C_2 as the cases where conditions induce high losses in crops caused by rust (IRR ≥ 7.18%), while C_3 cluster presents the cases with low risk of losses in crops (IRR < 7.18%).

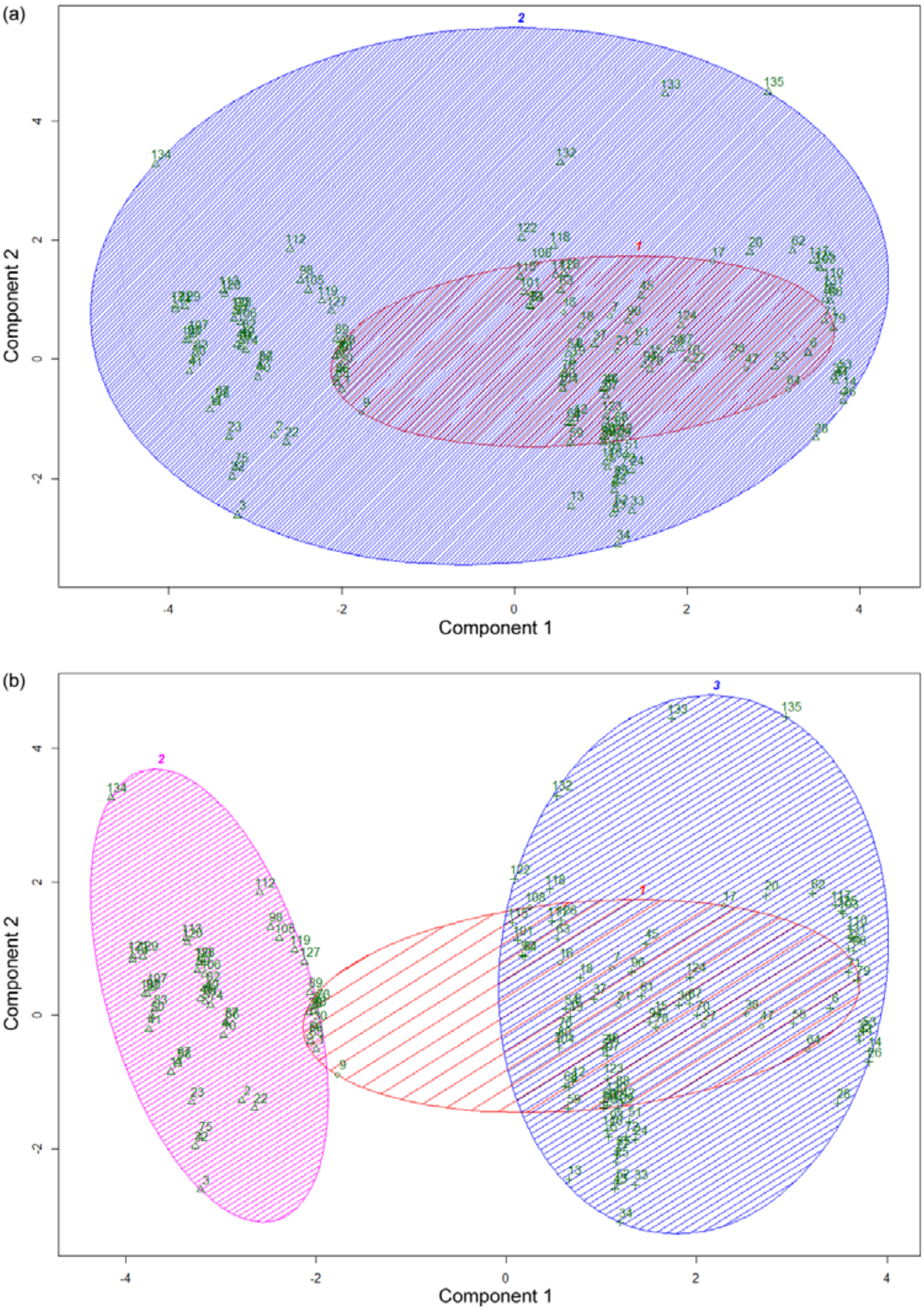
Figure 4 C4.5 decision tree for clusters generated by k-means


Coffee rust control in the last month
Si
No
Hours of relative humidity > 90% last month
<= 349
> 349
C1 (29)
C3 (5)
Hours of relative humidity > 90% last month
<= 341
> 341
C2 (1)
Crop age (months)
<= 48
> 48
C3 (48)
Crop age (months)
<= 50
> 50
C3 (4)
C2 (51)

3.1.4 Building the clusters

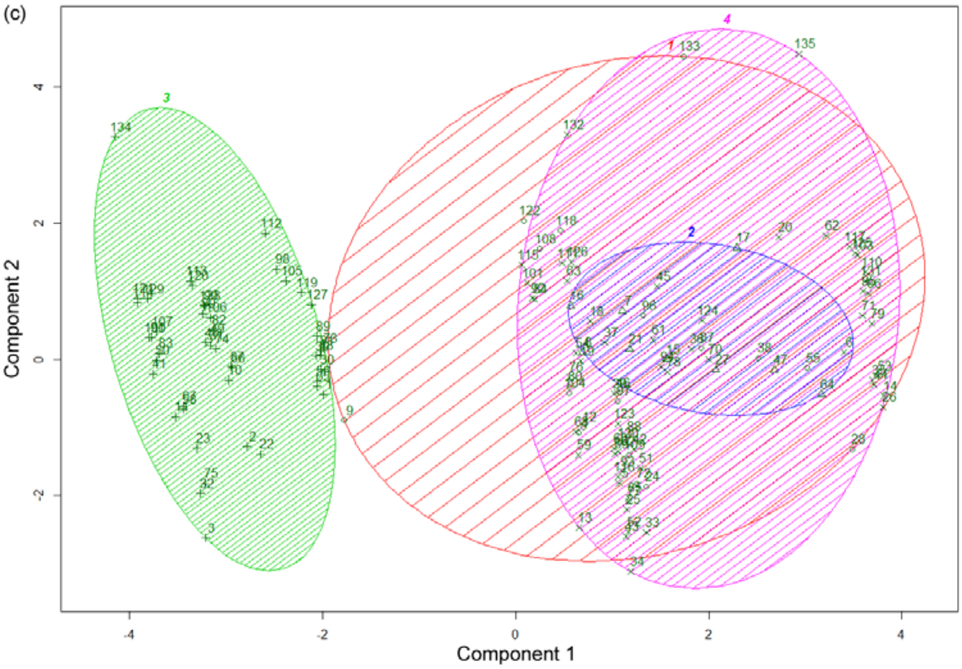
To choose the k suitable, the k-means algorithm was tested with $k = 2$, 3, 4 and the Euclidean distance. The clusters are displayed with CLUSPLOT graphical library, which creates a bivariate plot visualising a partition (clustering) of the data (Corrales et al., 2015c). All observation (instances) are represented by points in the plot, using principal components analysis (relative to the first two principal components) (Corrales et al., 2015b) available in CLUSPOT. The clusters $\{C_1, C_2,\ldots,C_n\}$; $k = n$ are again represented as ellipses, which are based on the average and the covariance matrix of each cluster; and their size is such that they contain all the points of their cluster (Figure 5).

Figure 5 K-means algorithm with *k* = 2, 3, 4 displayed with CLUSPLOT (see online version for colours)



Source: Authors own elaboration

Figure 5 K-means algorithm with $k = 2, 3, 4$ displayed with CLUSPLOT (continued) (see online version for colours)



Source: Authors own elaboration

The Figure 5(a) presents the k-means algorithm with $k = 2$, where the points of C_1 (red colour) has misclassified in correspondence to C_2 (blue colour), in this manner we cannot differentiate the C_1 from the C_2. On the other hand, $k = 3$ [Figure 5(b)], C_2 (pink colour) and C_3 (blue colour) are completely distinct. However, most of the points of C_1 (red colour) were misclassified as C_3. Lastly, $k = 4$ [Figure 5(c)], C_1 (red colour), C_2 (blue colour) and C_4 (pink colour) are overlapped, while the C_3 (green colour) differs of the other clusters.

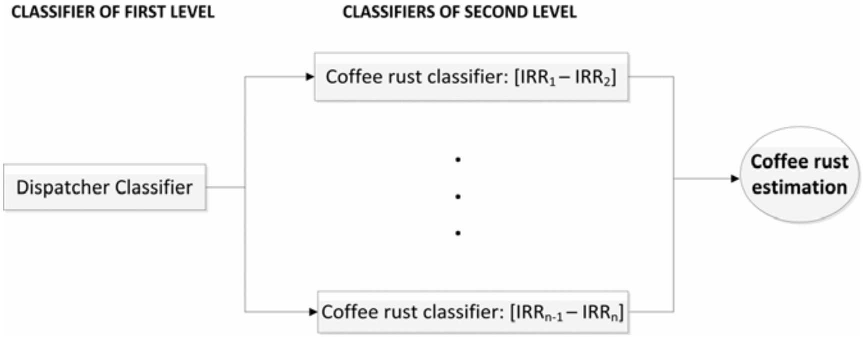
In this sense, we define the k-means algorithm with $k = 3$ [Figure 5(b)], because its clusters C_2 (pink colour) and C_3 (blue colour) are completely distinct. C_1 (red colour) has misclassified points due to its belonging to C_2 and C_3, besides the incoherent instances found by Bayesian network and C.4 decision tree. For these reasons, the observations of C_1 were deleted (29 instances) to avoid the noise. The new dataset has 109 instances, 52 of C_2 and 57 of C_3.

Once the final clusters (C_2 and C_3) have been defined, we are able to create three base classifiers for construction of the ensemble approach for coffee rust estimation which are explained in Section 3.2.

3.2 Ensemble approach for coffee rust infection estimation

Ensemble approach for coffee rust infection estimation is constructed based on data analysis module with two-level classifiers ensembles: the first level contains a BPNN named dispatcher classifier responsible for deciding which classifier of second level will estimate the IRR. The second level contains expert classifiers for solving a specific task. In this case we define two classifiers: the first classifier of second level (Figure 2 as coffee rust classifier: 0%–7.18%) is used to train a SVR with instances of C_3 with aim to detect the IRR among 0% and 7.18%. The second classifier of second level (Figure 2 coffee rust classifier: 7.18%–15%) trains a regression tree (M5) with instances of C_2 to classify the IRR among 7.18% and 15%. The number of classifiers of second level change according to cluster analysis and size of dataset as we can see in Figure 6; of this manner we can get *n* classifiers of second level where each classifier detects the IRR among [IRR_{n-1} – IRR_n]. Our processed dataset contains 109 instances with IRR values among 0% and 15%.

Figure 6 Two-level classifiers ensembles for coffee rust estimation


CLASSIFIER OF FIRST LEVEL
CLASSIFIERS OF SECOND LEVEL
Coffee rust classifier: [IRR_1 – IRR_2]
Dispatcher Classifier
Coffee rust estimation
Coffee rust classifier: [IRR_{n-1} – IRR_n]

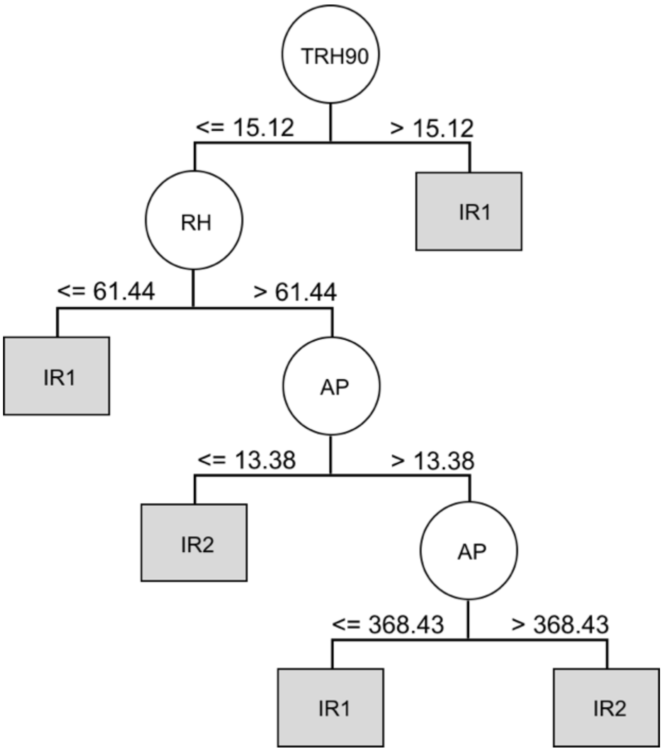
In this way, the final user types the inputs (weather conditions, physic crop properties, and crop management) and the ensemble approach returns the estimation of infection rate of rust.

3.3 Simple classifier for coffee rust growth estimation

To build the simple classifier, the target variable was obtained from the disease behaviour analysis, according to the impact value of coffee rust existing in the dataset between consecutive months. Thus, the infection rate is calculated by evaluating the increase or decrease in the incidence among the analyzed month and next month, getting two classes: *IR1 (≤ 0)*: reduction or latency, to negative or none infection rates; *IR2 (> 0)*: disease growth, to positive infection rates. From these considerations, it was constructed a decision tree from the C4.5 algorithm, as can be seen in Figure 7.

The generated DT accounted three attributes: temperature in hours with relative humidity > 90% (TRH90), mean relative humidity (RH), and daily accumulated precipitation (AP). Low temperature values during periods of leaf wetting is the prime determinant in the tree classification process. Also, relative humidity and rain present on crops lead to increase or decrease likelihood of infection, as might be established in the study of expert knowledge for this disease.

Figure 7 DT for coffee rust growth estimation



4 Experimental results

This section reports a number of experiments carried out to select the base classifiers used to detect the coffee rust incidence rate and the disease growth. Here we compare the results obtained by the ensemble approach against classical approaches as: simple classifiers and ensemble methods. With the dataset introduced in Section 2.2, we used a 10-fold cross validation to estimate the scores reported in the following figure and tables.

4.1 Selection of base classifiers

The families of supervised learning algorithms assessed were: SVMs, neural networks, Bayesian networks, decision trees, and *K* nearest neighbours. The selection criteria of these classifiers are based on previous surveys which show the suitable learners for classification and predictions tasks (Bhavsar and Ganatra, 2012), especially in the detection of crop diseases.

4.1.1 Dispatcher classifier

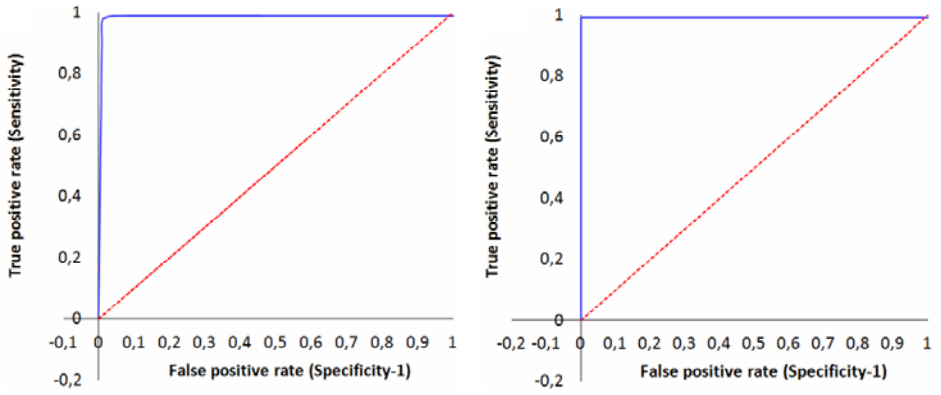
We tested the most relevant algorithms of supervised learning for classification tasks as SVM, BPNN, Naive Bayes (NB), C4.5 decision tree, and k nearest neighbours (K-NN)

(Corrales et al., 2015d) to choose the dispatcher classifier, computing precision, recall and F-measure as seen in Table 3.

Table 3 Precision, recall and F-measure for SVM, ANN, BN, DT and K-NN

Measures	*Supervised learning algorithms*				
	SVM	*BPNN*	*NB*	*C4.5*	*K-NN*
Precision	99.3%	99.3%	88.6%	96.3%	97.9%
Recall	99.3%	99.3%	87.4%	96.3%	97.8%
F-measure	99.3%	99.3%	87.5%	96.3%	97.8%

Figure 8 ROC curves for (a) SVM (b) BPNN (see online version for colours)



(a) (b)

The supervised learning algorithms in Table 3 present values greater than 88.6% for precision, 87.4% for recall and 87.5% for F-measure, which indicates low false positives and good outcomes to classify new instances. However, these measures are insufficient to choose the dispatcher classifier considering that SVM and BPNN have the same values (precision, recall and F-measure 99.3%). For this reason we used the ROC curve for evaluating the SVM and BPNN such as shows in Figure 8.

In this regard, the curves showed a good performance of SVM [Figure 8(a)] and BPNN [Figure 8(b)] to classify correctly new instances, with an area under a curve of 99.1% for SVM and 100% of BPNN. Based on the above, the BPNN was selected as dispatcher classifier.

4.1.2 Simple classifier for coffee rust growth estimation

In a similar way to dispatcher classifier selection, we tested some supervised learning algorithms for classification tasks as SVM, BPNN, NB and DT C4.5. As a result, we obtained some performance measures as precision, recall, and F-measure, which can be seen in Table 4.

Table 4 shows how the algorithm used to generate the simple classifier for estimating disease growth presents the best performance measures compared to the other supervised learning algorithms. Thus, C4.5 algorithm is better able to avoid the noise in the training set and classifies more relevant instances.

Table 4 Performance measures for DT classifier

Measures	*Supervised learning algorithms*			
	SVM	*BPNN*	*NB*	*C4.5*
Precision	91.5%	91.5%	77.9%	92.2%
Recall	91%	91%	77.7%	91.6%
F-measure	90%	90.9%	77.7%	91.5%

4.1.3 Classifiers of second level

We tested the main learning algorithms for prediction tasks as SVR, multilayer perceptron (MP), radial basis function network (RBFN), K nearest neighbours regression (K-NN R) and regression tree (M5) (Corrales et al., 2015d) to choose the classifiers of second level through Pearson's correlation coefficient (PCC), mean absolute error (MAE) and root mean squared error (RMSE) as seen in Table 5 and Table 6.

Table 5 Comparison: SVR, MP, RBF, K-NN R, M5 for selection of *1st-classifier-2nd* level

Measures	*Supervised learning algorithms*				
	SVR	*MP*	*RBF*	*K-NN R*	*M5*
PCC	0.43	0.39	-0.19	0.32	0.51
MAE	3.17%	3.62%	3.50%	3.22%	1.83%
RMSE	3.81%	4.83%	4.36%	4.04%	2.16%

In Table 5, the M5 regression tree is approaching a directly proportional relation (PCC = 0.51) among IRR detected (IRRP) and incidence rate of rust real (IRRR). In that manner M5 presents the least difference among IRRP and IRRR with MAE = 1.83% and RMSE = 2.16%. In accordance with the above, the M5 regression tree was selected as first classifier of second level (1st-classifier-2nd level).

In Table 6, K-NN R and SVR have highest value for positive correlations (PCC = 0.37 and PCC = 0.36 respectively), however, SVR presents the least difference among IRRP and IRRR (MAE = 1.20% and RMSE = 1.73%) regard to K-NN R (MAE = 1.31% and RMSE = 1.86%). Based on the above, SVR was selected as second classifier of second level (2nd-classifier-2nd level).

Table 6 Comparison: SVR, MP, RBF, K-NN R, M5 for selection of 2nd-classifier-2nd level

Measures	*Supervised learning algorithms*				
	SVR	*MP*	*RBF*	*K-NN R*	*M5*
PCC	0.36	0.32	0.27	0.37	0.08
MAE	1.20%	1.22%	1.24%	1.31%	1.49%
RMSE	1.73%	1.91%	1.74%	1.86%	2.08%

4.2 Evaluation of the ensemble approach for coffee rust estimation

This section presents the results obtained by the ensemble approach against simple classifiers and classical ensemble methods.

4.2.1 Ensemble approach vs. simple classifiers

Table 7 compares the outcomes obtained by the ensemble approach against simple classifiers as SVR, BPNN and regression tree (M5) which were tested in (Corrales et al., 2014) with the same dataset.

Table 7 Comparison of ensemble approach and simple classifiers

Measures	*Supervised learning algorithms*				
	Ensemble approach		*Simple classifiers*		
	1st-classifier-2nd level	*2nd-classifier-2nd level*	*SVR*	*BPNN*	*M5*
PCC	0.51	0.36	0.29	0.35	0.22
MAE	1.83%	1.20%	2.28%	2.34%	2.55%
RMSE	2.16%	1.73%	3.38%	3.31%	3.50%

The outcomes obtained by ensemble approach are better than simple classifiers; especially for *2nd-classifier-2nd level* where the instances are closer to each other regard to instances of *1st-classifier-2nd level* (Figure 5 for $k = 3$).

4.2.2 Ensemble approach vs. classical ensemble methods

Table 8 compares the outcomes obtained by the ensemble approach against classical ensemble methods as bagging, random subspaces, rotation forest and stacking. The classical ensemble methods as bagging used M5 as base classifier, random subspaces: K-NN R, rotation forest: M5, stacking three base classifiers: BPNN, K-NN R, M5 and SVR as meta-learner. We choose the four ensemble methods as the best outcomes to use the dataset explained in Section 2.2.

Table 8 Comparison of ensemble approach and classical ensemble methods

Measures	*Supervised learning algorithms*					
	Ensemble approach		*Classical ensemble methods*			
	1st-classifier-2nd level	*2nd-classifier-2nd level*	*Bagging*	*Ran. subspaces*	*Rot. forest*	*Stacking*
PCC	0.51	0.36	0.27	0.25	0.24	0.14
MAE	1.83%	1.20%	2.38%	2.38%	2.43%	2.41%
RMSE	2.16%	1.73%	3.34%	3.52%	3.37%	3.43%

The outcomes obtained by ensemble approach are better than classical ensemble methods. Bagging is the ensemble method with better results; nevertheless, simple classifiers as BPNN (PCC = 0.35; RMSE = 3.31%) and SVR (MAE = 2.28%) outperformed the results obtained by bagging (PCC = 0.27; MAE = 2.38%; RMSE = 3.34%).

5 Conclusions and future work

This paper presented an ensemble approach for coffee rust infection and growth estimation in Colombian crops. Using a simple classifier to estimate the coffee rust growth trend, combined with ensemble methods for disease infection estimation in crops, allows the characterisation of favourable conditions that influence it in terms of its evolution (current status and trend). Similarly, the consideration of expert knowledge in coffee rust during the training set definition results in a set of variables closely related to the disease, which are the main input for classifier development. Our approach outperformed the classical approaches as simple classifiers and ensemble methods in terms of PCC (0.51 of 1st-classifier-2nd level and 0.36 of 2nd-classifier-2nd level respect to 0.35 of BPNN and 0.27 of bagging), MAE (1.83% of 1st-classifier-2nd level and 1.20% of 2nd-classifier-2nd level respect to 2.28% of SVR and 2.38% of bagging) and RMSE (2.16% of 1st-classifier-2nd level and 1.73% of 2nd-classifier-2nd level respect to 3.31% of BPNN and 3.34% of bagging) which use the same dataset of coffee rust. The limitation encountered during this study was the absence of data from actual coffee crop. Especially in rust incidence rate samples due to the expensive collection process that requires big efforts in money and time. Accordingly, the results obtained on this study are not very precise.

In future studies we will try to tackle the insufficient data using different approaches such as synthetic data and incremental learning, which update the hypothesis of classifier using new individual data instances, without having to re-process past instances (Schlimmer and Granger, 1986). Also we will propose the use of weather time series data which are automatically capture each five minutes in weather station. We will analyze its behaviour with the rust infection rate.

Acknowledgements

The authors would like to thank the Telematics Engineering Group (GIT) of the University of Cauca and the Control Learning and Systems Optimization Group (CAOS) of the Carlos III University of Madrid, and KNIME® Analytics Platform, for the technical support. Finally, this work has been partially supported by AgroCloud project of the RICCLISA Program and the Spanish Goverment (under projects TRA2011-29454-C03-03 and TRA2015-63708-R), and Colciencias for PhD scholarship granted to MsC David Camilo Corrales.

References

Araujo, B.S. (2006) *Aprendizaje automático: conceptos básicos y avanzados: aspectos prácticos utilizando el software Weka*, Pearson Prentice Hall, España.

Avelino, J., Cristancho, M., Georgiou, S., Imbach, P., Aguilar, L., Bornemann, G., Läderach, P., Anzueto, F., Hruska, A.J. and Morales, C. (2015) 'The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions', *Food Secur.*, Vol. 7, No. 2, pp.303–321, doi:10.1007/s12571-015-0446-9.

Becker, S. (1977)' Diurnal periodicity in spore dispersal of Hemileia vastatrix in relation to weather factors', *Z. Pflanzenkrankh. Pflanzenschutz*, Vol. 84, No. 1, pp.577–591.

Bhavsar, H. and Ganatra, A. (2012) 'A comparative study of training algorithms for supervised machine learning', *Int. J. Soft Comput. Eng. IJSCE*, Vol. 2, No. 4, pp.74–81.

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

Corrales, D.C., Figueroa, A., Ledezma, A. and Corrales, J.C. (2015a) 'An empirical multi-classifier for coffee rust detection in Colombian crops', *15th International Conference on Computational Science and Its Applications – ICCSA 2015*, Banff, AB, Canada, 22–25 June 2015, Proceedings, Part I, Springer International Publishing, pp.60–74.

Corrales, D.C., Ledezma, A. and Corrales, J.C. (2015b) 'A conceptual framework for data quality in knowledge discovery tasks: a proposal', *J. Comput.*, Vol. 10, No. 6, pp.396–405.

Corrales, D.C., Ledezma, A. and Corrales, J.C. (2015c) 'A survey of data quality issues in knowledge discovery tasks', *VII Congreso Iberoamericano de Telemática – Workshop TIC-@agro*.

Corrales, D.C., Corrales, J.C. and Figueroa-Casas, A. (2015d) 'Towards detecting crop diseases and pest by supervised learning', *Ingeniería y Universidad*, Vol. 19, No. 1, pp.207–228.

Corrales, D.C., Ledezma, A., Peña, A., Hoyos, J., Figueroa, A. and Corrales, J.C. (2014) 'A new dataset for coffee rust detection in Colombian crops base on classifiers', *Sist. Telemática*, Vol. 12, No. 29, pp.9–23.

Ghosh, J. (2002) 'Multiclassifier systems: back to the future', *Proceedings of the Third International Workshop on Multiple Classifier Systems*, Springer-Verlag, 744269, pp.1–15.

Grubbs, F. (1969) 'Procedures for detecting outlying observations in samples', *Technometrics*, Vol. 11, No. 1, pp.1–21, doi:10.1080/00401706.1969.10490657.

Jain, A.K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognit. Lett.*, Vol. 31, No. 8, pp.651–666.

Kushalappa, A.C., Akutsu, M. and Ludwig, A. (1983) 'Application of survival ratio for monocyclic process of Hemileia vastatrix in predicting coffee rust infection rates', *Phytopathology*, Vol. 73, No. 1, pp.96–103.

Lasso, E., Thamada, T.T., Meira, C.A.A. and Corrales, J.C. (2015) 'Graph patterns as representation of rules extracted from decision trees for coffee rust detection', in Garoufallou, E., Hartley, R.J. and Gaitanou, P. (Eds.): *Metadata and Semantics Research, Communications in Computer and Information Science*, pp.405–414, Springer International Publishing [online] https://link.springer.com/chapter/10.1007/978-3-319-24129-6_35.

Li, L., Zou, B., Hu, Q., Wu, X. and Yu, D. (2013) 'Dynamic classifier ensemble using classification confidence', *Neurocomputing*, Vol. 99, pp.581–591, doi:10.1016/j.neucom.2012.07.026 [online] http://www.sciencedirect.com/science/article/pii/S092523121200608X.

Luaces, O., Rodrigues, L.H.A., Alves Meira, C.A. and Bahamonde, A. (2011) 'Using nondeterministic learners to alert on coffee rust disease', *Expert Syst. Appl.*, Vol. 38, No. 11, pp.14276–14283, doi:10.1016/j.eswa.2011.05.003.

Luaces, O., Rodrigues, L.H.A., Meira, C.A.A., Quevedo, J.R. and Bahamonde, A. (2010) 'Viability of an alarm predictor for coffee rust disease using interval regression', *Proceedings of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems* – Springer-Verlag, 1945889, Vol. Part II, pp. 337–346.

McAlister, D. (1879) 'The law of the geometric mean', *Proc. R. Soc. Lond.*, Vol. 29, pp.376–367, doi:10.1098/rspl.1879.0061.

Meira, C.A.A., Rodrigues, L.H.A. and de Moraes, S.A. (2009) 'Modelos de alerta para o controle da ferrugem-do-cafeeiro em lavouras com alta carga pendente', *Pesqui. Agropecuária Bras.*, Vol. 44, No. 3, pp.233–242.

Meira, C.A.A., Rodrigues, L.H.A. and Moraes, S.A. (2008) 'Analysis of coffee leaf rust epidemics with decision tree', *Trop. Plant Pathol.*, Vol. 33, pp.114–124, doi:10.1590/S1982-56762008000200005.

Mucherino, A., Papajorgji, P. and Pardalos, P. (2009) 'Clustering by k-means', in Du, D-Z. (Ed.): *Data Mining in Agriculture*, pp. 47–56, Springer, New York.

Nutman, F.J., Roberts, F.M. and Clarke, R.T. (1963) 'Studies on the biology of Hemileia vastatrix', *Berk. & Br. Trans. Br. Mycol. Soc.*, Vol. 46, No. 1, pp.27–44.

Pérez-Ariza, C.B., Nicholson, A.E. and Flores, M.J. (2012) 'Prediction of coffee rust disease using Bayesian networks', in Andrés Cano, M.G-O. and Nielsen, T.D. (Ed.): *The Sixth European Workshop on Probabilistic Graphical Models, DECSAI*, University of Granada.

Quinlan, J.R. (1993) *C4. 5: Programming for Machine Learning*, Morgan Kauffmann, San Francisco, CA, USA.

Ranawana, R. and Palade, V. (2006) 'Multi-classifier systems: review and a roadmap for developers', *Int. J. Hybrid Intell. Syst.*, Vol. 3, No. 1, pp.35–61.

Rivillas, C., Serna, C., Cristancho, M. and Gaitán, A. (2011) *Roya del Cafeto en Colombia: Impacto, Manejo y Costos del Control*, Bol. Téc., Chinchiná

Rivillas-Osorio, C., Serna-Giraldo, C., Cristancho-Ardila, M. and Gaitán-Bustamante, A. (2011) *La roya del cafeto en Colombia, impacto, manejo y costos de control, Avances Tecnicos Cenicafe. Cenicafé*, Chinchiná – Caldas – Colombia.

Schlimmer, J.C. and Granger Jr, R.H. (1986) 'Incremental learning from noisy data', *Mach. Learn.*, Vol. 1, No. 3, pp.317–354.

Sutton, J.C., Gillespie, T.J. and Hildebrand, P.D. (1984) 'Monitoring weather factors in relation to plant disease [Crop microclimate, electrical sensors, temperature and wetness gauges, sources of error]', *Plant Dis.*, Vol. 68, No. 1, pp.78–84.

Waller, J.M., Bigger, M. and Hillocks, R.J. (Eds.). (2007) *Coffee Pests, Diseases and their Management*, CAB International, Egham, Surrey, UK.