A Guideline for Building Large Coffee Rust Samples Applying Machine Learning Methods

Jhonn Pablo Rodríguez(✉), Edwar Javier Girón, David Camilo Corrales, and Juan Carlos Corrales

Department of Telematics Engineering, Engineering Telematics Group, University of Cauca, Popayán, Colombia
{jhonnpablo,edwardgb,dcorrales,jcorral}@unicauca.edu.co
http://www.unicauca.edu.co

Abstract. Coffee rust has become a serious concern for many coffee farmers and manufacturers. The American Phytopathological Society discusses its importance saying this: "the most economically important coffee disease in the world," while "in monetary value, coffee is the most important agricultural product in international trade". The early detection has inspired researchers to apply supervised learning algorithms on predicting the disease appearance. However, the main drawback of the related works is the few data samples of the dependent variable: Incidence Rate of Rust, since the datasets do not have a reliable representation of the disease, which will generate inaccurate classifiers. This paper provides a guide to increase coffee rust samples applying machine learning methods through a systematic review about coffee rust in order to select appropriate algorithms to increase rust samples.

Keywords: Synthetic data · Incidence · Crops · Dataset

1 Introduction

Coffee rust has become a serious concern for many coffee farmers and manufacturers. The American Phytopathological Society discusses its importance saying this: "the most economically important coffee disease in the world," while "in monetary value, coffee is the most important agricultural product in international trade". Without a solution, the effects on the coffee industry may soon be reflected in price and availability [1].

For several years, the disease was managed through the combination of various techniques such as quarantine, cultural management, fungicides and resistant crops. Due to the effectiveness of chemical control and the relatively limited damage caused by the disease, particularly at high altitudes, Mesoamerican coffee farmers and technical authorities considered it manageable. This view prevailed until the epidemic between 2008 and 2013 along Mesoamerica, from Colombia to Mexico, including Peru, Ecuador and some Caribbean countries [2]. Coffee farmers were desperate to obtain an answer to this terrible situation since the intensity was higher than anything previously observed, affecting a large number of countries including: Colombia, from 2008 to 2011, affecting an average of 31% of coffee production compared with the production in 2007; Central

P. Angelov et al. (eds.), *Advances in Information and Communication Technologies for Adapting Agriculture to Climate Change*, Advances in Intelligent Systems and Computing 687, https://doi.org/10.1007/978-3-319-70187-5_8

America and Mexico, in 2012–13, affecting an average of 16% of the production in 2013 compared with 2011–12 and an average of 10% in 2013–14 compared with 2012–13; and Peru and Ecuador in 2013 [2]. More specifically, in 2013, the Guatemalan government and the Guatemalan Nation Coffee agency declared a national state of emergency after a projection of nearly 15% crop loss in their region. The devastation has continued to spread due to higher temperatures in this region, which are making fungus growth at higher altitudes possible [3]. Higher temperatures may be linked to climate change. And several/many experts are worried about the persistence of these conditions (high temperatures) will not change in the near future. In this regard, several reports and experts proposed solutions related with early detection of the disease and the eradication of infected plants.

The early detection has inspired researchers to apply supervised learning algorithms on predicting the disease appearance. The data collected about conditions and soil fertility properties, physical properties and management of a coffee crop, can be used to forecast the rust infection rate. In the same way, weather conditions such as the minimum and maximum levels of temperature, humidity and rainy days can help to estimate the behavior of the disease. Several Colombian and Brazilian researches in supervised learning attempt to detect the incidence rate of rust (IRR) in coffee crops using Neural Networks, Decision Trees, Support Vector Machines, Bayesian Networks, K Nearest Neighbor, and Ensemble Methods [4–8]. However, the main drawback of the related works is the few data samples of the dependent variable: Incidence Rate of Rust, since the datasets do not have a reliable representation of the disease, which will generate inaccurate classifiers [5].

This paper provides a guide to increase coffee rust samples applying machine learning methods through a systematic review about coffee rust in order to select appropriate algorithms to increase rust samples. The paper is structured as follows: in Sect. 2, we describe the coffee rust disease and supervised learning concepts. Section 3 exposes the supervised learning approaches applied to coffee rust detection and the main challenges due to low accuracy of rust detecting models; Sect. 4 shows a systematic review of the approaches to generate synthetic data. Section 5, proposes a guideline for building large dataset of coffee rust based on systematic review of Sect. 4. Finally, the Sect. 6 presents the conclusions.

2 Background

2.1 Coffee Rust

Coffee rust is caused by the fungus Hemileia vastatrix, among the cultivated species C. Arabica is the most severely attacked. The disease causes defoliation, sometimes this one can lead to death of branches and crop losses. The first symptoms are small yellowish lesions that appear on the underside of the leaves, where the fungus has penetrated through the stomata. These lesions grow, coalesce and produce uredospores with their distinctive orange color. Chlorotic spots can be observed on the upper surface of the leaves. During the last stage of the disease, lesions become necrotic [2]. The progression of coffee rust depends on four factors that appear simultaneously [9]:

- **The host:** There are varieties of coffee plants susceptible and resistant to rust. Varieties such as Típica, Borbón and Caturra suffer severe rust attacks, while Colombia and Castillo varieties are highly resistant to rust.
- **Pathogenic organism:** Hemileia vastatrix lifecycle begins with the germination of uredospores in 2–4 h in optimal conditions. Within 24–48 h, infection is completed. Once the infection is completed, the underside of the leaf is colonized and sporulation will occur through the stomata [10].
- **Weather conditions:** Weather with constant precipitations, mainly in the afternoon and night with cloudy sky, high humidity in the plants and low temperatures are relevant factors for germination of rust. Spread of disease and its development is usually limited to the rainy season, while in dry periods the rust incidence is very low.
- **Agronomic practices:** This practice refers to properties of crop sowing (plant spacing, percentage of shade, etc.), application of fungicides and fertilizations on coffee crops with the aim to avoid several rust attacks.

In Colombia the incidence of rust is measured through the methodology developed by Centro Nacional de Investigaciones de Café (Cenicafé) [9], which is explained as follows:

Incidence Rate of Rust (IRR) is calculated for a plot with area lower or equal of one hectare. The methodology is composed by three steps:

1. The farmer must be standing in the middle of the first furrow, 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. It is worth mentioning that the same number of trees must be selected in every furrow (e.g. if the plot has 30 furrows, the farmer selects two coffee trees for each furrow).
3. Once the step 1 and 2 are finished, the Leaves of the selected Coffee Trees (LCT) are added as well as the Infected Leaves by Rust (ILR). Then, the Incidence Rate of Rust (IRR) must be computed using the following formula:

$$IRR = \frac{ILR}{LCT} \times 100 \tag{1}$$

Furthermore in Brazil, Procafé Foundation [11] proposes a methodology that enables measuring the incidence of rust through the following steps:

1. The farmer selects a random region of 4 m^2 from the plot.
2. From the region selected in the previous step, the farmer selects two coffee trees, located in different furrows, one in front of the other.
3. The trees are divided into three levels according to its height: high, medium, and low; and the branches are divided in quartiles by its size. One branch of the middle zone of the plant is chosen, then two leaves of that area are taken from the third or fourth part of the branch.

Finally, for each plot 25 regions are selected, from which 50 plants are chosen, therefore 100 leaves are collected. The samples are taken the first two days of each month.

2.2 Supervised Learning Techniques

Based on [4], in this section the main supervised learning algorithms, from coffee rust domain are explained:

A supervised learning process is based on the iteration of a training process from a dataset named training data. The training data consist of a set of examples. Each example is represented by a pair (x_i, y_i), where x_i is an attributes vector and y_i is the desired output value (also called class) of the example. A SL algorithm analyzes the training data and produces an inferred function which is called classifier (if y_i is discrete) or regression function (if y_i is continuous) [12]. In Table 1 an example of a dataset with three attributes is presented: Number of Days of Precipitation (NDP), average Daily Nighttime relative Humidity (DNH), average Daily Minimum Temperatures (DMT), which can take different values (discrete or continuous); and the desired output value [4].

Table 1. Four sample for training dataset

Attribute			Output value
NDP	DNH	DMT	IRR
2	96.1	14	65.23%
3	93.8	16	62.54%
4	95.7	15	57.32%
1	98.2	14	61.12%

The most commonly APPLIED algorithms are Decision Trees (DT), Bayesian Networks (BN), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and K Nearest Neighbor (KNN) [4].

3 Supervised Learning Techniques in Coffee Rust Detection

This systematic review took into account the inclusion criterion: Colombian and Brazilian research using supervised learning algorithms. And as an exclusion criterion: investigations not greater than 8 years. Systematic review was based on the following research question:

Are there researches that address the coffee rust by supervised learning?

18 papers were found (2008–2016) from 4 sources of information: IEEE Xplore (1 paper), ScienceDirect (1 paper), Springer Link (6 paper) and Google Scholar (10 papers). We defined 2 search queries: “coffee rust prediction” and “coffee rust detection”.

The papers found in the systematic review, contain approaches for coffee rust detection in Colombian and Brazilian crops, using supervised learning algorithms. These researches are detailed below:

3.1 Supervised Learning for Coffee Rust Detection in Colombian Crops

The dataset built in the Colombian researches was obtained for 18 plots from experimental farm Los Naranjos (21° 35′08′′ N, 76° 32′53′′ W), of the Supracafé enterprise, located in Cajibío (Cauca) [26]. The samples were recollected among years 2011–2013, with 147 instances and 21 attributes: 6 of weather conditions, 5 soil fertility properties, 6 physic crop properties, 4 crop management; the class represents the Incidence Rate of Rust (IRR). This dataset was used in several works presented in Table 2:

Table 2. Related works for coffee rust detection in Colombian crops

Work	Year	Algorithm
[13]	2014	ANN, SVM, RT
[5]	2015	ANN, SVM, RT
[14]	2016	Two-level classifier ensembles using Back Propagation Neural Networks, Regression Tree M5 and Support Vector Regression
[15]	2014	SVM
[16]	2015	DT

Briefly, Colombian researchers using SVM (4 papers), ANN (2 papers) and DT/RT (4 paper), but the most used are SVM and ANN, since this class of algorithms deliver results accurate to the end user, these are less prone to overfitting than other methods, tolerance to the noise, Accuracy in general, Tolerance to irrelevant attributes, easy to understand and speed in its learning and classification [4].

3.2 Supervised Learning for Coffee Rust Detection in Brazilian Crops

Brazilian researchers built a dataset from physic crop properties and weather conditions. These data were collected in the experimental farm Procafé (South latitude 21° 34′00′′ longitude West 45° 24′22′′ and altitude 940 m) located in Varginha, Minas Gerais, during the years 1998–2011 [17]. The final dataset includes 182 instances. Table 3 are presented the works related with the Brazilian dataset:

Thus, the algorithms used in Brazil are DT, SVM, SVR, ANN, RF and BN, focusing in its priority to generate an easy interpretation model based on graphs like the DT and generate accurate results with SVM.

Table 3. Related works for coffee rust detection in Brazilian crops

Work	Year	Algorithm
[17]	2008	DT
[18]	2012	BN
[19]	2013	ANN, DT, Random Forest (RF), SVM
[8]	2015	Ensembles Methods with SVM, ANN, DT
[20]	2009	DT
[6]	2011	DT
[21]	2014	SVM, ANN, DT, RF
[22]	2011	SVR
[23]	2010	SVR

3.3 Discussion

The algorithms used in the last years are Support Vector Machines (SVM), Decision Trees (DT), Bayesian Networks (BN), Nearest Neighbor (KNN) and Artificial Neural Networks (ANN), but the algorithms SVM and ANN are the most used for the precision in the categorization of results, however the algorithms present deficiencies in their interpretation, since classifiers built by these algorithms do not generate a visual representation, in contrast to the DT and algorithms the BN, which allow the user to observe the classifier through a representation based on graphs [4]. On the other hand, algorithms such as: KNN and BN are good by their speed of learning (training phase) [24–26].

In accordance with the system review, Brazil and Colombia are the countries that address the coffee rust detection through supervised learning. However, the researchers found are limited due lack of data in measures of Infection Rate of Rust, due to the high costs and time invested for the collection of data rust infection. As a result, datasets cannot represent faithfully the total population, generating low accuracy in the results obtained by classifiers [5]. Section 4 describes approaches for resolving this kind of problem, through the generation of synthetic data.

4 Synthetic Data Generation

The few amount of samples do not let to the models to represent important characteristics of the population, therefore the models constructed are affected in its precision [5]. For this systematic review, the inclusion criterion was taken into account: research to increase the number of samples in a dataset. And as an exclusion criterion: proposed researches that do not have a benchmark analysis with traditional algorithms. Systematic review was based on the following research question:

Which are synthetic data approaches most used for lack of data?

We found 26 papers (2000–2015), considering 5 search queries: “Synthetic Data Generation”, “Imbalanced Dataset”, “Over-Sampling”, “Virtual Sample Generation” and “Interpolation Algorithm”, it was found from 4 sources of information: IEEE Xplore

(10 papers), ScienceDirect (5 papers), Springer Link (7 papers) and Google Scholar (4 papers).

From the systematic review conducted previously, were found 4 approaches to addressing the lack of data and their respective algorithms for synthetic data generation as shown in Table 4:

Table 4. Approaches and synthetic data algorithms

Approach	Algorithms	Description	Works
Over-Sampling	MDO, SMOTE, ADASYN, RWO SAMPLING, BORDERLINE SMOTE2 MSMOTE, BORDERLINE SMOTE1, BORDERLINE SMOTE, C-SMOTE, SMOTE-I, DSMOTE	Creates new synthetic instances for the minority class	[27–39]
Interpolation	Stair, S-Spline, Bicubic, Lanczos, Nearest Neighbor, Fractals, Linear	Determinates the values of a function at positions lying between its samples	[40–44]
Classifiers	AdaBoost, RAMOBoost, AdaBoost.M2, AdaBoost.M1, DataBoost.IM, SMOTEBoost, DataBoost, SMOTE Bagging, OverBagging	Classification or regression model that aims to predict the value of output variable from certain independent input variables	[30, 35, 36, 45–48]
Copy of Data	Bagging, Regression Trees, Random Forest, Statistical Distributions	Creates a copy from original data with a different representation that not reveal private information	[49–51]

To construct the guideline for increasing coffee rust samples, we used the approaches: oversampling, interpolation, and classifiers. Copy of data is not used because for this kind of problem is necessary use the original representation of data.

5 A Guideline for Building Large Coffee Rust Samples

In this section, we propose a guide to increase coffee rust samples. Figure 1 presents the process for generate discrete or numeric rust samples.

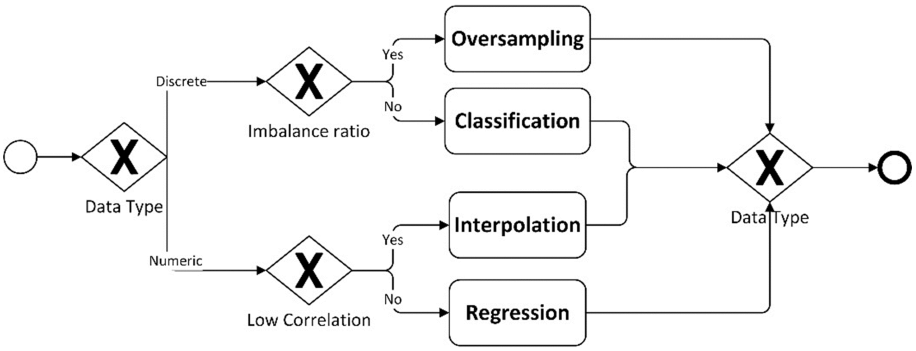

Fig. 1. Guide for increase coffee rust samples

When the coffee rust samples are discrete, the imbalance ratio must be evaluated. Commonly the Imbalance Ratio (IR) is used to measure the distribution of the classes:

$$IR = \frac{C^+}{C^-} \tag{2}$$

Where C+ represents the size of the majority class and C− the size of the minority class [52]. *Oversampling* techniques are used if IR > 1; otherwise *Classification* techniques are applied. The techniques for increase the discrete coffee rust samples are presented below.

5.1 Oversampling

Oversampling is used to increase examples from minority classes with aim have equitable distribution of the classes [28]. The algorithm most used are called SMOTE: Synthetic Minority Over-sampling Technique [27]; RUS Boost [55], Balance Cascade [56] and Easy Ensemble [56] also are considered Oversampling algorithms. Other algorithms used are MDO [35], SMOTE [31, 33, 39], ADASYN [29, 30], RWO SAMPLING [37], BORDERLINE SMOTE2 [38], MSMOTE [36], BORDERLINE SMOTE1 [38], BORDERLINE SMOTE [38], C-SMOTE [32], SMOTE-I [57], DSMOTE [34].

5.2 Classification

Classification algorithms are efficient methods to increase discrete samples where each new value is obtained from related cases in the whole set of records. Besides the capability to increase the coffee rust samples with plausible values that are as close as possible to the true value, classification algorithms should preserve the original data structure and avoid to distort the distribution of the original samples. The algorithms based in neighbours are the most used [58]. The algorithm most used are AdaBoost [30, 35, 36, 45], RAMOBoost [30], AdaBoost.M2 [36], AdaBoost.M1 [36], DataBoost.IM [59], SMOTEBoost [30, 36, 46], DataBoost [48], SMOTE Bagging [36], OverBagging [36].

In case of the coffee rust samples are numeric, a correlation analysis must be done. The range of values for the correlation coefficient is −1 to 1. A correlation of −1 indicates a perfect negative correlation, while a correlation of 1 indicates a perfect positive correlation. Values close to 0 indicates a low correlation. A regression approach is used when the dependent variables have a relationship strong with the predictor variables (correlation coefficient close to 1 or −1) [53]; otherwise the interpolation approach is used (correlation coefficient close 0) because these methods are focused in mathematic functions [54]. The techniques for increase the numeric coffee rust samples are presented below.

5.3 Interpolation

Interpolation is the process of determining the values of a function at positions lying between its samples. It achieves this process by fitting a continuous function through the numeric input samples. The interpolation can be addressed of two ways: univariate and multivariate [60].

Univariate Interpolation
The univariate interpolation is defined for values of a function f(x) = y, where only take part two variables (x, y), in which x is the series with full data and y is the incompleteness variable in the serie. Inside univariate interpolation, there are algorithms that interpolate the y values, of which three obtain good results: lineal interpolation, K-nearest neighborhood (KNN), and cubic spline interpolation [61–64].

Multivariate Interpolation
The univariate interpolation fitted of two-dimensional data points, while the multivariate interpolation make the points fit finding the surface that provides an exact fit to a series of multidimensional data points, considering a series of N distinct dimensional data points (x1, y1), (x2, y2), … (xN, yN), where $Xi = \left(X_i^1, X_i^2, \ \ldots \ X_i^d\right)$ is a vector for each i = 1, 2, … N. By this interpolation we find a function f: $R^d \rightarrow R$ such that: f(x1) = y1, f(x2) = y2, … f(xN) = yN [65]. The algorithms most used are Inverse Distance Weighting (IDW) and Kriging [66].

5.4 Regression

Similarly, to Classification algorithms, the coffee rust is treated as dependent variable and a regression is performed to increase coffee rust samples. Regression analysis is a machine learning approach that aims predict the value of continuous output variables (coffee rust samples) from certain independent input variables (e.g. temperature, humidity, etc.), via automatic estimation of their latent relationship from data [67]. Linear regression, logistic regression [68], regression trees [69], support vector regression [70] and multi-layer perceptron [71] are typical choices.

6 Conclusions

In this paper, a guideline for increasing the coffee rust samples was made. The first step to create the guideline was to make a systematic review about coffee rust in order to select appropriate algorithms to increase rust samples.

When the guideline for increasing the coffee rust samples is used, we consider important follow the next observations [72]:

- *Distance between meteorological station and coffee trees:* if a weather station is away from a coffee plot, the weather measurements are inaccurate, because coffee plot can have micro-climate influenced by: coffee plot orography and properties of crop sowing such as: plant spacing, shade on coffee trees, etc. Unfortunately, the weather stations are very expensive to have one per coffee plot.
- *Information about application of fungicides on coffee plots:* if fungicides are applied on coffee plots before germination of rust, the weather conditions can not be relevant factors to increase the rust incidence. We consider necessary this information to build a correct regression model based on meteorological variables.
- *Consider a margin of error in IRR measurements:* the insufficient data due to the expensive collection process that requires large expenditures of money and time [7]. The farmers must select 3 branches for each 60 coffee trees (minimum) per plot [23]. Usually one plot have 10000, 5000, or 2500 coffee trees [1], given that the maximum number of IRR measures that we can obtain for one plot are: 0.6%, 1.2%, or 2.4% respectively. Besides a coffee farm has over one coffee plot.
- Due to the coffee rust is an element of a agronomic pathosystem, it is necessary to study how the factors about crop administration (fungicides, production levels, and so on), the meteorological factors (humidity, temperature and so on) and the fungus development interact with each other, and how this interaction contribute to increase the coffee rust disease.

7 Acknowledgments

We thank to the Telematics Engineering Group (GIT) of the University of Cauca for the technical support. Finally, this work has been partially supported by AgroCloud project of the RICCLISA Program, and Colciencias for PhD scholarship granted to MsC. David Camilo Corrales and Cauca Innovacción Nucleus Group for the Young investigator program to Ing. Edwar Javier Girón.

References

1. Arneson, P.A.: Coffee rust. Plant Health Instr. (2000)
2. Avelino, J., et al.: The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions. Food Secur. **7**(2), 303–321 (2015)
3. A Solution to the Coffee Rust Epidemic: How Spectrophotometry May Provide the Answers. HunterLab Horizons Blog, 12 January 2015

4. Corrales, D.C., Corrales, J.C., Figueroa-Casas, A.: Towards detecting crop diseases and pest by supervised learning. Ing. Univ. **19**(1), 207–228 (2015)
5. Corrales, D.C., Figueroa, A., Ledezma, A., Corrales, J.C.: An empirical multi-classifier for coffee rust detection in colombian crops. In: Gervasi, O., Murgante, B., Misra, S., Gavrilova, M.L., Rocha, A.M.A.C., Torre, C., Taniar, D., Apduhan, B.O. (eds.) Computational Science and Its Applications, ICCSA 2015, pp. 60–74. Springer, Heidelberg (2015)
6. Cintra, M.E., Meira, C.A.A., Monard, M.C., Camargo, H.A., Rodrigues, L.H.A.: The use of fuzzy decision trees for coffee rust warning in Brazilian crops. In: 2011 11th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 1347–1352 (2011)
7. Cesare di Girolamo, L.H.R.: Potencial de técnicas de mineração de dados para modelos de alerta da ferrugem do cafeeiro (2013)
8. Thamada, T.T., Rodrigues, L.H.A., Meira, C.A.A.: Predição da taxa de progresso da ferrugem do cafeeiro por meio de ensembles. Predicting infection rate of coffee rust by ensembles (2015)
9. Rivillas Osorio, C.A.: La roya del cafeto en Colombia, impacto, manejo y costos de control. Cenicafé: Chinchiná - Caldas - Colombia (2011)
10. Nutman, F.J., Roberts, F.M., Clarke, R.T.: Studies on the biology of Hemileia vastatrix Berk. & Br. Trans. Br. Mycol. Soc. **46**(1), 27–44 (1963)
11. Garcia, A.L.A.: RESUMO METODOLÓGICO DE AVALIAÇÃO DAS VARIÁVEIS FENOLÓGICAS E FITOSSÂNITÁRIAS DO SISTEMA DE AVISOS FITOSSÂNITÁRIOS DO MAPA/PROCAFÉ, Varginha, Brasil (2011)
12. Ng, A.: CS 229 machine learning course materials. In: Supervised learning. University of Stanford (2003)
13. Corrales, D.C., Ledezma, A., Andrés, J.P.Q., Hoyos, J., Figueroa, A., Corrales, J.C.: A new dataset for coffee rust detection in Colombian crops base on classifiers. Sist. Telemática **12**(29), 9–23 (2014)
14. Corrales, D.C., Casas, A.F., Ledezma, A., Corrales, J.C.: Two-level classifier ensembles for coffee rust estimation in colombian crops. Int. J. Agric. Environ. Inf. Syst. **7**, 41–59
15. Corrales, D.C., Peña, A.J.: Early warning system for coffee rust disease based on error correcting output codes: a proposal. Rev. Ing. Univ. Medellín **13**(25), 59–64 (2014)
16. Lasso, E., Thamada, T.T., Meira, C.A.A., Corrales, J.C.: Graph patterns as representation of rules extracted from decision trees for coffee rust detection. In: Garoufallou, E., Hartley, R.J., Gaitanou, P. (eds.) Metadata and Semantics Research, pp. 405–414. Springer, Heidelberg (2015)
17. Meira, C.A.A., Rodrigues, L.H.A., Moraes, S.A.: Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. Trop. Plant Pathol. **33**(2), 114–124 (2008)
18. Pérez-Ariza, C.B., Nicholson, A.E., Flores, M.J.: Prediction of coffee rust disease using Bayesian networks. In: Andrés Cano, M.G.-O., Nielsen, T.D. (eds.) The Sixth European Workshop on Probabilistic Graphical Models. University of Granada, Granada, Spain (2012)
19. Cesare di Girolamo, L.H.R.: Desenvolvimento e seleção de modelos de alerta para a ferrugem do cafeeiro em anos de alta carga pendente de frutos (2013)
20. Meira, C.A.A., Rodrigues, L.H.A., de Moraes, S.A.: Warning models for coffee rust control in growing areas with large fruit load. Pesqui. Agropecuária Bras. **44**(3), 233–242 (2009)
21. di Girolamo Neto, C., Rodrigues, L.H.A., Meira, C.A.A.: Modelos de predição da ferrugem do cafeeiro (Hemileia vastatrix Berkeley & Broome) por técnicas de mineração de dados, 22 2014. http://www.alice.cnptia.embrapa.br/handle/doc/991078. Accessed 3 Feb 2016
22. Luaces, O., Rodrigues, L.H.A., Alves Meira, C.A., Bahamonde, A.: Using nondeterministic learners to alert on coffee rust disease. Expert Syst. Appl. **38**(11), 14276–14283 (2011)

23. Luaces, O., Rodrigues, L.H.A., Meira, C.A.A., Quevedo, J.R., Bahamonde, A.: Viability of an alarm predictor for coffee rust disease using interval regression. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds.) Trends in Applied Intelligent Systems, pp. 337–346. Springer, Heidelberg (2010)
24. Bhavsar, H., Ganatra, A.: A Comparative Study of Training Algorithms for Supervised Machine Learning
25. "Supervised Machine Learning: A Review of Classification …," 11:38:43 UTC
26. Segrera Francia, S., Moreno García, M.N.: Multiclasificadores: métodos y arquitecturas, March 2006. http://gredos.usal.es/jspui/handle/10366/21727. Accessed 29 Dec2015
27. Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 853–867. Springer, New york (2005)
28. He, H., Ma, Y.: Foundations of imbalanced learning. In: Imbalanced Learning: Foundations, Algorithms, and Applications, p. 216. Wiley-IEEE Press (2013)
29. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
30. Thanathamathee, P., Lursinsap, C.: Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. Pattern Recognit. Lett. **34**(12), 1339–1347 (2013)
31. Wong, G.Y., Leung, F.H.F., Ling, S.H.: A novel evolutionary preprocessing method based on over-sampling and under-sampling for imbalanced datasets. In: 39th Annual Conference of the IEEE Industrial Electronics Society, IECON 2013, pp. 2354–2359 (2013)
32. He, G., Han, H., Wang, W.: An over-sampling expert system for learning from imbalanced data sets. In: 2005 International Conference on Neural Networks and Brain, ICNN B 2005, vol. 1, pp. 537–541 (2005)
33. Pengfei, J., Chunkai, Z., Zhenyu, H.: A new sampling approach for classification of imbalanced data sets with high density. In: 2014 International Conference on Big Data and Smart Computing (BIGCOMP), pp. 217–222 (2014)
34. Mahmoudi, S., Moradi, P., Akhlaghian, F., Moradi, R.: Diversity and separable metrics in over-sampling technique for imbalanced data classification. In: 2014 4th International eConference on Computer and Knowledge Engineering (ICCKE), pp. 152–158 (2014)
35. Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans. Knowl. Data Eng. **28**(1), 238–251 (2016)
36. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(4), 463–484 (2012)
37. Zhang, H., Li, M.: RWO-Sampling: a random walk over-sampling approach to imbalanced data classification. Inf. Fusion **20**, 99–116 (2014)
38. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) Advances in Intelligent Computing, pp. 878–887. Springer, Heidelberg (2005)
39. Kerdprasop, N., Kerdprasop, K.: Predicting rare classes of primary tumors with over-sampling techniques. In: Kim, T., Adeli, H., Cuzzocrea, A., Arslan, T., Zhang, Y., Ma, J., Chung, K., Mariyam, S., Canción, X. (eds.) Database Theory and Application, Bio-science and Bio-technology, pp. 151–160. Springer, Heidelberg (2011)
40. Malpica, J.A.: Splines interpolation in high resolution satellite imagery. In: Bebis, G., Boyle, R., Koracin, D., Parvin, B. (eds.) Advances in Visual Computing, pp. 562–570. Springer, Heidelberg (2005)

41. Hung, K.-W., Siu, W.-C.: Learning-based image interpolation via robust k-NN searching for coherent AR parameters estimation. J. Vis. Commun. Image Represent. **31**, 305–311 (2015)
42. Rui, L., Qiong, L.: Image sharpening algorithm based on a variety of interpolation methods. In: 2012 International Conference on Image Analysis and Signal Processing (IASP), pp. 1–4 (2012)
43. Bentbib, A.H., El Guide, M., Jbilou, K., Reichel, L.: A global Lanczos method for image restoration. J. Comput. Appl. Math.
44. Shi, Z., Yao, S., Li, B., Cao, Q.: A novel image interpolation technique based on fractal theory. In: 2008 International Conference on Computer Science and Information Technology, ICCSIT 2008, pp. 472–475 (2008)
45. Sun, Y., Kamel, M.S., Wang, Y.: Boosting for learning multiple classes with imbalanced class distribution. In: 2006 Sixth International Conference on Data Mining, ICDM 2006, pp. 592–602 (2006)
46. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) Knowledge Discovery in Databases. PKDD 2003, pp. 107–119. Springer, Heidelberg (2003)
47. Viktor, H.L., Guo, H.: Multiple classifier prediction improvements against imbalanced datasets through added synthetic examples. In: Fred, A., Caelli, T.M., Duin, R.P.W., Campilho, A.C., de Ridder, D. (eds.) Structural, Syntactic, and Statistical Pattern Recognition, pp. 974–982. Springer, Heidelberg (2004)
48. Guo, H., Viktor, H.L.: Boosting with data generation: improving the classification of hard to learn examples. In: Orchard, B., Yang, C., Ali, M. (eds.) Innovations in Applied Artificial Intelligence, pp. 1082–1091. Springer, Heidelberg (2004)
49. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Comput. Stat. Data Anal. **55**(12), 3232–3243 (2011)
50. Anderson, J.W., Kennedy, K.E., Ngo, L.B., Luckow, A., Apon, A.W.: Synthetic data generation for the internet of things. In: 2014 IEEE International Conference on Big Data (Big Data), pp. 171–176 (2014)
51. Albuquerque, G., Lowe, T., Magnor, M.: Synthetic generation of high-dimensional datasets. IEEE Trans. Vis. Comput. Graph. **17**(12), 2317–2324 (2011)
52. Verbiest, N., Ramentol, E., Cornelis, C., Herrera, F.: Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced classification data. In: Advances in Artificial Intelligence, IBERAMIA 2012, pp. 169–178 (2012)
53. Törn, A.A.: Correlation coefficients of linear regression models of human decision making. Omega **8**(3), 393–394 (1980)
54. Field, A., Miles, J., Field, Z.: Discovering Statistics Using R (2012)
55. Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A.: RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man Cybern. Part Syst. Hum. **40**(1), 185–197 (2010)
56. Albayrak, A.S.S.: Alleviating the Class Imbalance problem in Data Mining (2013)
57. SMOTE: Synthetic Minority Over-sampling Technique. https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/chawla2002.html. Accessed 19 June 2017
58. Beretta, L., Santaniello, A.: Nearest neighbor imputation algorithms: a critical evaluation. BMC Med. Inform. Decis. Mak. **16**(Suppl), 3 (2016)
59. Guo, H., Viktor, H.L.: Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. SIGKDD Explor. Newsl. **6**(1), 30–39 (2004)

60. Mohanty, P.K., Reza, M., Kumar, P., Kumar, P.: Implementation of cubic spline interpolation on parallel skeleton using pipeline model on CPU-GPU cluster. In: 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 747–751 (2016)
61. Phillips, G.M.: Univariate interpolation. In: Interpolation and Approximation by Polynomials, pp. 1–48. Springer, New York (2003)
62. Keogh, E., Chu, S., Hart, D., Pazzani, M.: An online algorithm for segmenting time series. In: Proceedings 2001 IEEE International Conference on Data Mining, pp. 289–296 (2001)
63. Hamed, Y., Shafie, A., Mustaffa, Z.B., Idris, N.R.B.: An application of K-Nearest Neighbor interpolation on calibrating corrosion measurements collected by two non-destructive techniques. In: 2015 IEEE 3rd International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), pp. 1–5 (2015)
64. Li, H., Wan, X., Liang, Y., Gao, S.: Dynamic time warping based on cubic spline interpolation for time series data mining. In: 2014 IEEE International Conference on Data Mining Workshop, pp. 19–26 (2014)
65. Multivariate - Interpolation - Approximation - Maths Reference with Worked Examples. http://www.codecogs.com/library/maths/approximation/interpolation/multivariate.php. Accessed 20 Feb 2017
66. Influence of DEM interpolation methods in Drainage Analysis. https://www.researchgate.net/publication/237116945_Influence_of_DEM_interpolation_methods_in_Drainage_Analysis. Accessed 20 Feb 2017
67. Yang, L., Liu, S., Tsoka, S., Papageorgiou, L.G.: A regression tree approach using mathematical programming. Expert Syst. Appl. **78**, 347–357 (2017)
68. Magnani, M.: Techniques for Dealing with Missing Data in Knowledge Discovery Tasks (2004)
69. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Taylor & Francis (1984)
70. Vapnik, V., Golowich, S.E., Smola, A.J.: Support vector method for function approximation, regression estimation and signal processing. In: Mozer, M.C., Jordan, M.I., Petsche, T. (eds.) Advances in Neural Information Processing Systems 9, pp. 281–287. MIT Press (1997)
71. Neural Networks: A Comprehensive Foundation (2nd edn.) Neural Networks: A Comprehensive Foundation. ResearchGate. https://www.researchgate.net/publication/233784957_Neural_Networks_A_Comprehensive_Foundation_2nd_Edition_Neural_Networks_A_Comprehensive_Foundation. Accessed 16 June 2017
72. Corrales, D.C., Gutierrez, G., Rodriguez, J.P., Ledezma, A., Corrales, J.C.: Lack of data: is it enough estimating the coffee rust with meteorological time series? In: Computational Science and Its Applications, ICCSA 2017, pp. 3–16 (2017)