Application of classification-tree models to characterize the mycobiota of grapes on the basis of origin

Rita Serra[1], Anália Lourenço[2], Orlando Belo[2] and Armando Venâncio[1]

[1]Centro de Engenharia Biológica and [2]Departamento de Engenharia Informática, Universidade do Minho, Campus de Gualtar, Braga, Portugal

Summary

Classification-tree (CT) models are a simple and robust exploratory data analysis technique that can be used in classification, regressions and summaries of data. They distill complex ecological relationships into simplified rules and identify the species necessary for sample classification on the basis of detailed ecological inventories. The usefulness of this technique to characterize and represent differences in the grape mycobiota of distinct origins was evaluated.
Grapes from four Portuguese winemaking regions were selected for a 3-year study: Alentejo, Douro, Ribatejo and Vinhos Verdes. The mycobiota of grapes was assessed with plating methods and the frequencies of isolations of the fungal taxa identified in 32 samples were used as a training dataset.
The CT algorithm selected the fungal taxa and respective frequencies to classify grapes according to its region of origin. The ten-fold cross-validation technique was used for model evaluation. The success rate of the model was quantified and expressed in the number of correctly classified samples overall and into region. Furthermore, model refinement was performed using attribute selection algorithms and class redefinition. A simple tree model was generated that classified grapes into three regional origins: Douro, South (Alentejo and Ribatejo classes together) and Vinhos Verdes, on the basis of the incidence of *Aspergillus niger* aggregate and *Penicillium thomii* in samples with an accuracy of 82%. The merits and demerits of these models are discussed.

Key words Mycobiota , Wine, Grapes, *Aspergillus*, *Penicillium*, Classification-tree models

The mycobiota of grapes have been extensibly studied recently as fungi are responsible for several mycotoxin hazards in wine, especially ochratoxin A (OTA). The results of the work of several authors suggest that differences exist in the frequency and fungal species to which grapes are exposed in distinct winemaking regions [1-3,12-15]. Nevertheless, comparative studies have not been performed to study the variation in the mycobiota of grapes with the region of origin, independent of other factors such as annual variation. To characterize the differences of the mycobiota of grapes originating from distinct geographical origins, a 3-year study was conducted in 4 Portuguese winemaking regions. The mycobiota of 32 grape samples was assessed at harvest time, and the data were analysed with classification-tree (CT) models [4].

Corresponding address:
Armando Venâncio
Centro de Engenharia Biológica
Universidade do Minho
Campus de Gualtar
4710-057 Braga, Portugal
Tel.: +351 253 604 400
Fax:+351 253 678 986
E-mail: avenan@deb.uminho.pt

©2006 Revista Iberoamericana de Micología
Apdo. 699, E-48080 Bilbao (Spain)
1130-1406/01/10.00 €

Classification-tree (CT) models are a multivariate exploratory analysis technique used in classification, regression and summarization of data, for descriptive and predictive purposes. CTs are used to classify objects on the basis of one or more predictor variables. The tree consists of a set of decision rules, applied in a sequential manner, until each object has been assigned to a specific class. The first decision rule, applied at the 'root node' of the tree to the values of all objects along one or more predictor variables, has two possible outcomes: objects are either sent to a terminal node (leaf), which assigns a class, or to an intermediate node, which applies another decision rule. Ultimately, all objects are sent to a terminal node and assigned a class. In the simplest type of CT, the splits are binary (each parent node is attached to two daughter nodes) and the decision rules are univariate (based on a single variable). CTs can be based on continuous or discrete predictor variables, or a mixture of both (when univariate splits are used), and the trees are generally constructed by recursive partitioning (i.e. a given predictor variable can be used in more than one decision rule).

CT analysis offers some advantages over traditional statistic classification methods such as i) ability to handle data measured on different scales; ii) lack of any assumptions concerning the frequency distributions of the data in each of the classes; and iii) ability to handle non-linear relationships between features and classes. CTs are flexible data-driven tools and allow the development of a model the form of which is directly a function of a particular data set. They can be used for feature selection/

reduction and classification purposes. Unlike classification methods based on neural networks, CTs are a white box model, meaning the analyst can interpret a decision tree. The decision process is fully transparent, and its success is quantified by statistics associated with the models. As a result, they are excellent tools to describe and summarize data.

The models can be evaluated in relation to a test set or to an evaluation methodology such as cross-validation or bootstrapping, to interpret what has been observed and to generalise for future observations [16]. This performance evaluation can be useful to model refinement, which can be performed by attribute selection, to describe better the data and improve model accuracy in identifying unseen samples. This is of utmost importance in detecting problems such as overfitting, which happens when the model tries to "particularize" in an attempt to better describe the data rather than generalizing. The evaluation techniques are helpful to model refinement.

CTs are not new to ecological studies [9]. They have been used to synthesise ecological information gathered on different habitats without sacrificing ecological specificity [5]. They distill complex ecological relationships into highly simplified rules bases and identify only those indicator species necessary for habitat classification from detailed ecological inventories. Nevertheless, they are not commonly used in fungal ecology studies [7].

In this work, the purpose was to characterize the main differences in the fungal species of grapes according to their region of origin using CT models.

Materials and methods

Grape samples

Sampling. Grape samples were composed of 10 grape bunches collected according to two diagonal transects in the vineyard. Eleven Portuguese vineyards located in four winemaking regions were selected for the study: Douro (three vineyards), Vinhos Verdes (three vineyards), Alentejo (two vineyards) and Ribatejo (three vineyards). Portugal mainland is well described and is located between the parallels 36° 57' 39'' W and 42° 9' 8'' W (latitude North) and the meridians 6° 11' 10'' W and 9° 22' 5'' W (longitude West). Douro and Vinhos Verdes region are located in north Portugal, at latitudes 41° and 41° to 42° N, and longitudes 7° and 8°, respectively. Alentejo and Ribatejo are located in south Portugal, at latitudes 38° and 39° N, and longitudes 7 and 8°, respectively. The climate is Sub Mediterranean in Vinhos Verdes, and Mediterranean in the other regions mentioned, according to the Rivas-Martinez criteria [11]. The samples were collected in the vineyard between late August and September in the harvest seasons of 2001, 2002 and 2003, near to the harvest date decided by the producers. Grapes were from distinct grape varieties used for commercial winemaking, and the producers took all decisions regarding grape production. A total of 32 grape samples were collected in Alentejo (6), Douro (9), Ribatejo (9) and Vinhos Verdes (8).

Mycological analysis of grapes. The mycoflora of grapes were determined as described previously [14]: a total of 50 berries (five berries *per* bunch) of each sample were plated in Dichloran Rose Bengal Chloramphenicol medium (DRBC) (Oxoid, UK) and incubated at 25 °C in the dark for one week. *Aspergillus* and *Penicillium* were isolated and identified morphologically to species level. However, *Aspergillus niger* was referred to as *A. niger* aggregate because of the complexity of that taxon.

Classification modeling

Training dataset. The 32 samples were used with 62 predictive attributes. The attributes were all fungal taxa identified from samples (20 genera, 15 *Aspergillus* and 27 *Penicillium* species) and the class attribute was the region of origin. The number of colonized berries by each species in the sample was used as abundance, which varied from 0 to 50 (corresponding to 0% and 100% of colonization, respectively).

Classification algorithm. The training set was analysed using the J4.8 algorithm, which is a Java implementation [16] of the well-known C4.5 learning scheme [10]. The notion of entropy, introduced by Claude Shannon in Information Theory, was used to measure the informative value of the predictive attributes.

Output model. The output was a tree model, followed by several statistic measures and the "confusion" matrix. In the leaves of the model, the numbers in parentheses are (i) how many instances from the training set were classified by the node in that region, and (ii) the number of instances that were incorrectly classification by the node (if any, if just one figure is presented, then only correctly classified instances are presented). A confusion matrix contains information about actual and predicted classifications undertaken by the classifier. It is a square matrix that shows the various classifications and misclassifications of the model in a compact area. Each row in the confusion matrix represents an observed class, each column represents a predicted class, and each cell provides the number of samples in the intersection of those two classes.

Evaluation methodology. The obtained classification-tree models were then evaluated using the 10-fold cross-validation approach [8]. The dataset was divided into 10 subsets, ensuring that each class is represented with approximately equal proportions in all subsets. Then, each subset is used for testing and the remaining nine for training purposes. The error estimates are averaged to yield an overall estimate.

Attribute selection. In model refinement, WEKA CfsSubsetEval attribute selection algorithms were used and the methods selected were *BestFirst* and *RankSearch.*

Software. All classification modeling was performed using the Waikato Environment for Knowledge Analysis (WEKA) tool developed within the scope of the Waikato University Machine Learning project (http://www.cs.waikato.ac.nz/~ml/index.html). Weka is a comprehensive suite of state-of-the-art machine learning and data mining algorithms [6]. It is open-source platform-independent software that provides a user interface component to non-programmers.

Results

CT model based on the full mycobiota data of samples. The tree model obtained using all fungal taxa as predictive attributes (Figure 1) was able to correctly identified 29 out of the 32 grape samples (91%) with the full training dataset. Four fungal taxa were selected to classify grape origin: *A. niger* aggregate, *Botrytis* sp., *Penicillium corylophilum* and *Penicillium thomii*. *Botrytis* sp. being used twice with distinct criteria of abundance. There is one leaf to Douro and Vinhos Verdes classes and two leaves for Alentejo and Ribatejo classes. The performance of the model in the given classes is provided in table 1. The interpretation of the classification model in the classes mentioned is as follows: seven samples of berries from the

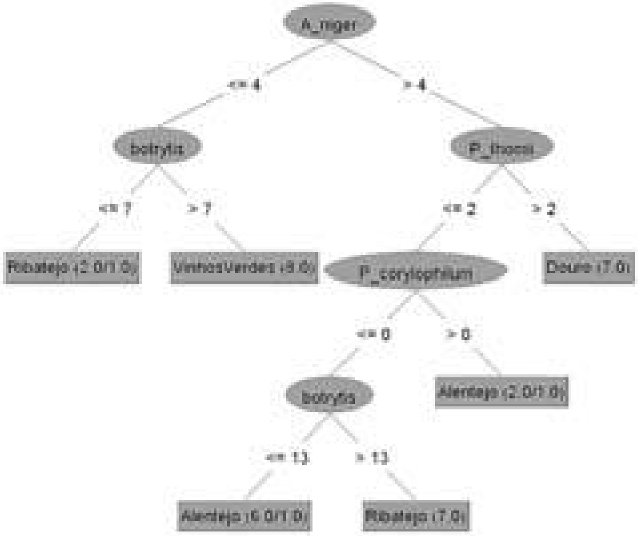

Figure 1. Classification tree output for describing region classes based on all attributes.

Table 1. Confusion matrix of the samples classified by the classification-tree model presented in figure 1.

Region	Alentejo	Douro	Ribatejo	Vinhos Verdes
Full training set				
Alentejo	**6**	0	0	0
Douro	1	**7**	1	0
Ribatejo	1	0	**8**	0
Vinhos Verdes	0	0	0	**8**
10-fold cross-validation				
Alentejo	**2**	0	3	1
Douro	0	**6**	1	2
Ribatejo	2	0	**5**	2
Vinhos Verdes	0	2	1	**5**

Each row in the confusion matrix represents the region of origin, each column represents a predicted region, and each cell counts the number of samples in the intersection of those two classes (the correctly identified samples are in bold).

Douro region were characterized by having more than 8% colonization with *A. niger* and more than 4% with *P. thomii*; all eight grape samples from the Vinhos Verdes region had less than 8% colonization of *A. niger* and more than 14% of *Botrytis* sp. colonization; six grape samples from Alentejo had more than 8% colonization of *A. niger* and less than 4% of *P. thomii*; two samples had the presence of *P. corylophilum*, one being from Alentejo; six samples had no *P. corylophilum* and less than 26% *Botrytis* sp. colonization, five being from Alentejo, while seven Ribatejo samples had more than 26% *Botrytis* sp. colonization. The remaining two grape samples had less than 8% colonization of *A. niger* and less than 14% of *Botrytis* colonization in the samples, one of which was from Ribatejo.

Using the evaluation 10-fold cross validation methodology, the CT model correctly identified 18 out of 32 samples (56% success). The number of correctly classified and missclassified samples in each class are indicated in table 1. The samples most successfully classified were those from North Portugal - Douro (67%) and Vinhos Verdes (62%).

Model refinement. In an attempt to improve the success of the model, it was refined in terms of attribute selection. The following fungal taxa were selected: *A. niger*, *P. expansum*, *P. thomii* and *Ulocladium* sp. The tree model (Figure 2) used as predictive attributes, *A. niger*, *P. thomii* and *Ulocladium* sp. Twenty-eight out of the 32 samples were correctly identified using the full training data set (= 88% success).

The model also selected *A. niger* and *P. thomii* as the most predictive attributes. The decision path to classify Douro samples was the same, but the decision path used to classify Vinhos Verdes was simplified, as the *Botrytis* node disappeared. The tree branch leading to Alentejo and Ribatejo samples starts with the split in the incidence of *Ulocladium* and uses *A. niger* incidence twice. Three leaves were generated to classify Alentejo samples and one leaf to classify Ribatejo samples. The success of this model with 10-cross fold-validation was higher (69% success). Douro and Vinhos Verdes samples were correctly classified with 78% and 88% success, respectively, but the classification of the South regions Alentejo and Ribatejo was less successful (Table 2). Fifty % of the Alentejo samples were incorrectly identified as originating from Ribatejo,

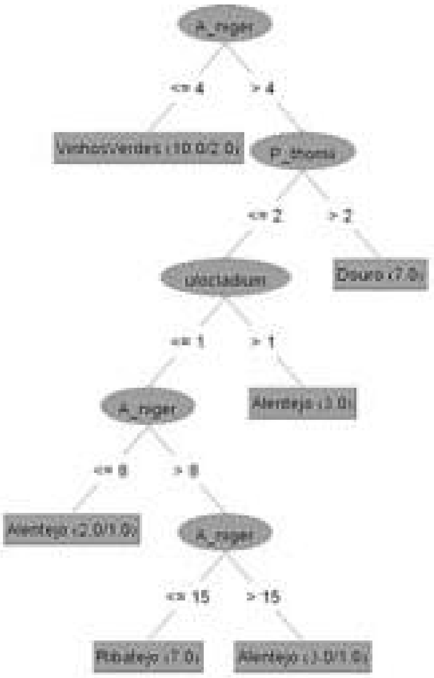

Figure 2. Classification tree output for describing region classes based on attribute selection.

Table 2. Confusion matrix of the samples classified by the classification-tree model presented in figure 2.

Region	Alentejo	Douro	Ribatejo	Vinhos Verdes
Full training set				
Alentejo	**6**	0	0	0
Douro	1	**7**	0	1
Ribatejo	1	0	**7**	1
Vinhos Verdes	0	0	0	**8**
10-fold cross-validation				
Alentejo	**3**	0	3	0
Douro	0	**7**	1	1
Ribatejo	2	1	**5**	1
Vinhos Verdes	0	1	0	**7**

Each row in the confusion matrix represents the region of origin, each column represents a predicted region, and each cell counts the number of samples in the intersection of those two classes (the correctly identified samples are in bold).

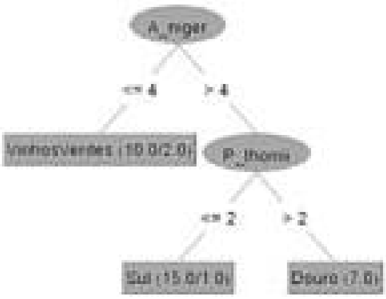

Figure 3. Classification tree output for describing three region classes (Douro, Vinhos Verdes and Sul [=South]) based on attribute selection.

Table 3. Confusion matrix of the samples classified by the classification-tree model presented in figure 3.

Region	South (Alentejo and Ribatejo)	Douro	Vinhos Verdes
Full training set			
South	**14**	0	1
Douro	1	**7**	1
Vinhos Verdes	0	0	**8**
10-fold cross-validation			
South	**13**	1	1
Douro	1	**7**	1
Vinhos Verdes	2	0	**6**

Each row in the confusion matrix represents the region of origin, each column represents a predicted region, and each cell counts the number of samples in the intersection of those two classes (the correctly identified samples are in bold).

while 22% of Ribatejo samples were incorrectly classified as originating from Alentejo, according to the model created with a 10-fold cross-validation. Due to the difficulty of the models to discriminate between Alentejo and Ribatejo samples, the two classes were merged and a new model was created to classify samples from Douro, Vinhos Verdes and South regions (Figure 3). Attribute selection was performed, and two predictive attributes were selected: *A. niger* and *P. thomii*.

The model was simpler than previous ones and classified 29 out of the 32 (91%) of the samples correctly in the three sample classes. Three leaves were created, one for each class. The following decision rules were generated: "where *A. niger* colonizes less than 8% of grape samples 10 observations occurred in the training data set and eight were from Vinhos Verdes; where *A. niger* colonizes more than 8% of the samples and *P. thomii* colonizes less than 4%, 15 observations occurred in the training data set and 14 were from South regions; where *A. niger* colonizes more than 8% of the samples and *P. thomii* colonizes more than 4%, seven observations occurred in the training data set and all were from Douro". Twenty-six out of the 32 samples (82%) were classified successfuly with a 10-fold cross-validation (Table 3).

Discussion

Classification tree modeling. CT models were able to classify grapes according to the origin with 91% success based on the full mycobiota of the samples. From the 62 attributes, the classification algorithm selected four fungal taxa as split variables with one *(Botrytis)* being used twice. The model used as the criterion the incidence of the selected species in the samples. The interpretability of the model was good. The samples of the four regions were characterized according the incidence of the taxa selected. The differences in the mycobiota between them were summarized and represented according to straight forward decision rules. Whereas the samples from Douro and Vinho Verdes were classified by one set of rules each, the samples from Alentejo and Ribatejo were classified with two sets of decision rules each, reflecting less homogeneity in the samples than those from Douro and Vinho Verdes. This was reflected in the lower performance of the model with the 10-fold cross-validation evaluation technique. The lower performance of the model with evaluation is an indicator of "over fitting". This happens when the classifier generalizes on particular aspects of data, unrepresentative of the whole class, resulting into good descriptive abilities but poor predictive accuracy [16]. This usually reflects a small sample size in the training dataset.

Prior to attribute selection, a new tree was generated. The model used *A. niger* and *P. thomii* as first splits, but used a variable not previously selected, *Ulocladium*, to discriminate between Alentejo and Ribatejo samples. Attribute selection improved the model accuracy, but the South samples from Alentejo and Ribatejo were frequently misclassified. The differences between the mycobiota of both regions were subtle, not allowing a clear distinction between the dataset and technique used. Nevertheless, the 82% success of the model to classify samples of Vinhos Verdes, Douro and Southern regions was considered satisfactory. The final model used the two first split attributes from previous models, *A. niger* and *P. thomii*. The variables in CT models are selected to create splits that maximize the resulting node homogeneity; therefore the variables used in early splits can be considered to be more important [9].

Variation in the grape mycobiota due to its region of origin. The results confirmed that the region of origin influences markedly the mycobiota to which grapes are exposed. The differences were mainly in the spoilage fungi, particularly *A. niger* and *P. thomii*. Quantitative differences were the most evident between grapes of distinct origins. Vinhos Verdes was easily characterized by the low incidence of *A. niger* in the samples, which is consistent with previous results [14]. This is of importance as *A. niger* is a species capable of producing the mycotoxin OTA that appears to be adapted to Mediterranean climates where it can be the dominant species of the grape mycobiota. The Mediterranean regions Douro, Ribatejo and Alentejo were discriminated by the incidence of *P. thomii* but no reason was found for the high incidence of this species in Douro samples. Nevertheless *P. thomii* is not directly relevant regarding mycotoxin production. With the training data set available and the data analysis technique used, it was found that the South samples (Alentejo and Ribatejo regions) share the same trend in the mycobiota to which are exposed.

Value of classification trees in comparative fungal studies. The ability of CTs to summarize and represent differences in the diversity of samples, both in qualitative and quantitative terms is very useful. The descriptive ability of classification trees is their main advantage for these studies. This technique facilitates the task of determining differences in the distribution of fungal taxa and indicator species of particular habitats or sites, thereby reducing the number of variables to be studied as has been shown in habitat assessment studies [5].

Nevertheless, as CTs are data driven tools, they are susceptible to over fitting due to small data sets. Another problem of using small data sets is the instability of the trees generated due to the effects of small variations on sample size. In our studies we confirmed this pitfall when we added new samples from new sites to the unrefined model (data not shown). Small data sets are a practical and unavoidable problem in detailed ecological studies. Nevertheless, over fitting can be detected with evaluation techniques and reduced with model refinement using attribute selection or model readjustment into new classes. Therefore, model evaluation by techniques such as 10-fold cross-validation is essential and can help in the model refinement process to guarantee the quality of the results. Furthermore this evaluation technique is adequate in small data sets as it does not require removing samples from the training data set for modeling.

In conclusion, these techniques can (a) prove the relevance of comparative studies of fungal diversity as exploratory techniques, (b) select indicator species of particular habitats or sites, (c) summarize differences between sample classes and (d) evaluate the generalizations made from small data sets.

The authors gratefully acknowledge the support of the EC, Quality of Life Programme (QoL), Key Action 1 (KA1) on Food, Nutrition and Health; contract number QLK1-CT-2001-01761-Wine-Ochra Risk. R. Serra and A. Lourenço were supported from Fundação para a Ciência e Tecnologia by grants SFRH/BPD/2827/2004 and SFRH/BD/8242/2002, respectively.

References

1. Battilani P, Pietri A, Bertuzzi T, Languasco L, Giorni P, Kozakiewicz Z. Occurrence of ochratoxin A-producing fungi in grapes grown in Italy. J Food Protect 2003; 66: 633-636.
2. Bau M, Bragulat MR, Abarca ML, Minguez S, Cabañes FJ. Ochratoxigenic species from Spanish wine grapes. International J Food Microbiol 2005; 98: 125-130.
3. Bellí N, Pardo E, Marín S, Farré G, Ramos AJ, Sanchis V. Occurrence of ochratoxin A and toxigenic potential of fungal isolates from Spanish grapes. J Sci Food Agricult 2004; 84: 541-546.
4. Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and decision trees. Belmont, Wadsworth, 1984.
5. Cohen MJ, Lane CR, Reiss KC, Surdick JA, Bardi E, Brown MT. Vegetation based classification trees for rapid assessment of isolated wetland condition. Ecological Indicators 2005; 5: 189-206.
6. Holmes G, Hall M. A Development environment for predictive modeling in foods. Int J Food Microbiol 2002; 73: 351-362.
7. Kenkel NC, Booth T. Multivariate analysis in fungal ecology. En: Carroll GC, Wicklow DT (Eds.) The fungal community: its organization role in the ecosystem. New York, Marcel Dekker, Inc., 1992: 209-227.
8. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. En: Mellish CS (Ed.) Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, Morgan Kaufmann, 1995: 1137-1143.
9. Miller J, Franklin J. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecol Model 2002; 157: 227-247.
10. Quinlan R. C4.5: Programs for machine learning. San Mateo, Morgan Kaufmann Publishers, 1993.
11. Rivas-Martinez S, Arregui JL. Bioclimatology of the Iberian Peninsula. Itinera Geobotanica 1999; 13: 41-47.
12. Rocha Rosa CA, Palácios V, Combina M, Fraga ME, Oliveira Rekson A, Magnoli CE, Dalcero AM. Potential ochratoxin A producers from wine grapes in Argentina and Brazil. Food Addit Contam 2002; 19: 408-414.
13. Sage L, Garon D, Seigle-Murandi F. Fungal microflora and ochratoxin A risk in french vineyards. J Agricult Food Chem 2004; 52: 5764-5768.
14. Serra R, Abrunhosa L, Kozakiewicz Z, Venâncio A. Black *Aspergillus* species as ochratoxin A producers in Portuguese wine grapes. Int J Food Microbiol 2003; 88: 63-68.
15. Serra R, Braga A, Venâncio A. Mycotoxin-producing and other fungi isolated from grapes for wine production, with particular emphasis on ochratoxin A. Res Microbiol 2005; 156: 515-521.
16. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques with Java implementations. San Mateo, Morgan Kaufmann Publishers, 1999.