Identifying the botanical origin of honeys with boosted decision trees

Laure-Anne Minsart, Marie Warnier, Etienne Bruneau

CARI ASBL, Place Croix du Sud 4, 1348 Louvain-la-Neuve, BELGIUM

Introduction

Determination of botanical origin of honey is an important step in order to provide value added to honey production, and especially in case of protected appellation.
By the past, number of methods have been tested to classify honeys into botanical types, using only sugar concentration, amino acids, pollens profile or organoleptic criteria. But none of these methods, taken separately, could provide a reliable botanical origin determination. Therefore, at present, this identification is the result of a synthesis of chemical, melissopalynologic and organoleptic specifications of a sample. However, even if this methodology is powerful, it requires great expertise and the outcome may be distorted due to human subjectivity.
In this context, we decided to develop a statistical decision support tool, using all these analytical variables.
The methodology chosen is a decision tree combined to boosting, called « Boosted Decision Tree », developed by Freund in 1999. A decision tree is a sequence of binary splits of the data. It is a powerful tool, but with a major disadvantage, its unstability. Indeed, a small change in the training data can produce a large change in the tree. This is remedied by the use of boosting, which builds up several trees, in order to add a confidence margin to the proposed types generated by the decision tree.
For our decision support tool, we decided to build up a tree based on the analytical results of all monofloral honeys analysed by CARI between 2009 and 2012. The Boosting decision Tree obtained was used to predict the botanical origin of honeys analysed at the beginning of 2013 season.

Material & Methods

Data used for the creation of the tree come from the analysis carried out by CARI laboratory on 406 monofloral honeys between 2009 and 2012. The botanical origin of each honey has been determined, at the time of analysis, by two experts, on the basis of analytical results.
306 variables, classified in 5 different category, were used for the construction of the tree :

- physico-chemical variables: conductivity, pH, equivalent pH, acidity, saccharase and diastase activity
- pollen variables : percentage of each pollen
- sugar variables : percentage of mono, di and trisaccharides
- organoleptic variables : odor, aroma and flavor descriptors
- visual variables : color

All qualitatives variables were transformed into binary variables.
The program which builds up the boosting tree was created with the programming language “R”, using the library “rpart” and “Adabag”. This last library was lightly modified in order to extract the importance of each variable in the construction of the tree. The boosting tree was created using a boosting of 50 iterations. 26 different botanical types were taken into account by the boosted decision tree (Table 1).

Table 1 : Botanical types of the boosted decision tree

Type	**Common name (English)**
Robinier faux-acacia	Robinia
Agrumes	Citrus
Baies roses	Pink Peppercorn
Bourdaine	Alder Buckthorn
Châtaignier	Chestnut
Châtaignier & ronces	Chestnut & Brambles
Colza	Rapeseed
Eucalyptus	Eucalyptus
Evodia	Evodia
Fruitiers	Fruit trees

Fruitiers & saule	Fruit trees & Willow
Lavande	Lavander
Litchi	Lychee
Miellat de résineux	Resinous honeydew
Palmier	Palm
Pissenlit	Dandelion
Retama	Retama
Romarin	Rosemary
Ronces	Brambles
Ronces & Trèfles	Brambles & Clover
Sarrasin	Buckweath
Saule	Willow
Thym	Thyme
Tilleul	Lime
Tournesol	Sunflower
Trèfles	Clover

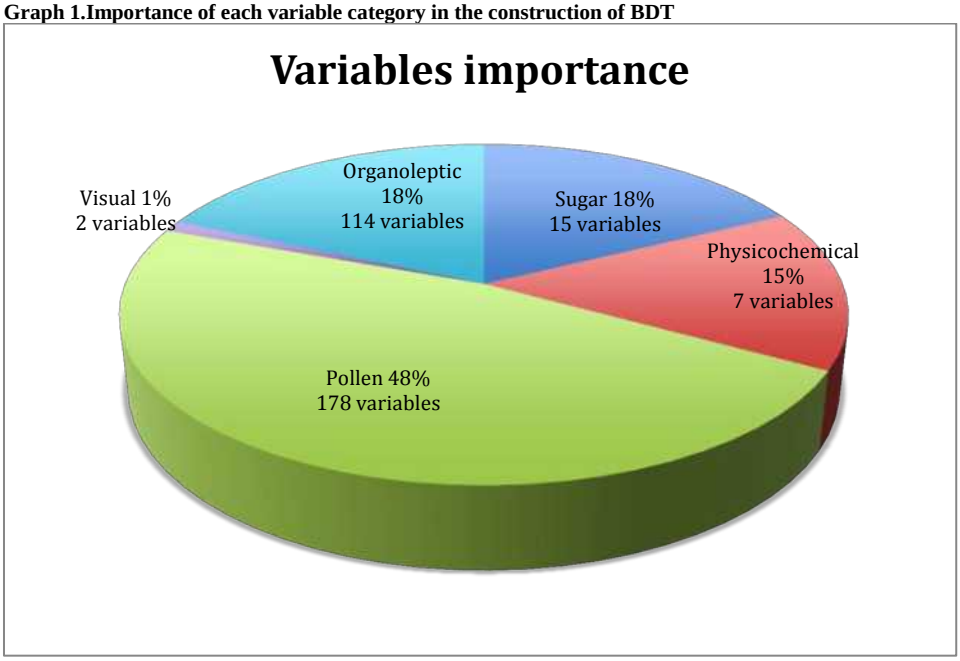
Once the boosting tree was built up, it was used to predict the botanical origin of honeys analysed at the beginning of 2013, following their analytical results.

Results

1) Boosted decision tree : honeys from 2009-2012

The boosted decision tree classifies the different honeys according to nodes and levels, linked to different analytical variables. It is thus important to carry out a variable analysis in order to understand which variable may have the greatest impact on the tree ramifications.
« R » program permits to extract the relative importance of each variable on the construction of the tree. Graph n°1 shows the relative importance between the different categories of variables on the construction of the decision tree.

Graph 1.Importance of each variable category in the construction of BDT



However, we have to pay attention on the fact that the percentages presented on this graph for each category are a sum of the individual percentages obtained for each variable present in this category.
Conductivity is for example the most important variable for the construction of the tree, as its importance rises up to 9,7%. Thus around 65% of the whole physicochemical category importance comes from conductivity. For sugars, glucose and fructose contents represent around 8,5% importance.

2) Prediction on 2013 honeys

Out of 75 honey analysis carried out by CARI laboratory in the beginning of 2013 season, 13 honeys were identified as monofloral samples, given the results obtained for the different analysis (physico-chemical, sugars, pollens, color and organoleptic) and following two experts' opinion. These 13 samples consisted in :

- 10 rapeseed honey
- 1 palm honey
- 1 rosemary honey
- 1 jujube honey

Analytical results of these 13 honeys were afterwards injected into the decision tree in order to predict their botanical origine. « R »program presents the outcome of each honey as a probability profile to belong to different origins, as shown in figure 2 for the sample n°14684.

Figure 2. Prediction profile for a Colza honey sample

```
individu  14684  , class  Colza  , bien classe = TRUE
-----------------------------------------------------------
type :  Colza  , proba:  0.614675107965911
type :  Tournesol  , proba:  0.12143056361069
type :  Agrumes  , proba:  0.0629587490456149
type :  Acacia  , proba:  0.0591173906844131
```

12 out of these 13 honeys were well predicted. This means that the major origin type obtained in the prediction profile corresponds well to the botanical origin (class) determined by the laboratory. Only the jujube honey, which was not yet caracterised in the tree, was therefore misclassified.

Discussion and perspectives

Boosted decision trees are a powerful tool when trying to classify data using a high number of variables. Furthermore, boosting prevents variations upon the programmation answer .
Results obtained for the prediction of botanical origin in honey are very promising. Using this tool, we could really get a decision support, in addition to the typical human analysis, for monofloral, but also for polyfloral honeys. Indeed, for the latter, the prediction profile could also help to find dominant nectars (results not shown). This program could definitely avoid subjectivity due to human factors.
However, in order to go a step further, the database for the creation of the boosted tree should be continuouly enriched, in order to get more precision in the answer. We should try also to determine a connection between the probability profile and the purity level of the sample. Indeed, it is of great importance to differentiate monofloral sample with samples only having a dominant character. Also, the program should be adapted to honeydew analysis.

Reference

- Alfaro-Cortes, E., M. Gamez-Martinez et N. Garcia-Rubio (2013). Package adabag : Applies multiclass AdaBoost.M1, AdaBoost-SAMME and Bagging, Version 3.1. Rap. tech.
- CRAN. Anklam, E. (1998). A review of the analytical methods to determine the geographical and botanical origin of honey. Food chemistry 63.4, 549–562.
- Azeredo, L. d. C., M. Azeredo, S. De Souza et V. Dutra (2003). Protein contents and physicochemical properties in honey samples of Apis mellifera of different floral origins. Food Chemistry 80.2, 249–254. Bertin-Mahieux, T.
- Cordella, C. B., J. S. Militao, M.-C. Cle´ment et D. Cabrol-Bass (2003). Honey characterization and adulteration detection by pattern recognition applied on HPAEC- PAD profiles. 1. Honey floral species characterization. Journal of agricultural and food chemistry 51.11, 3234–3242.
- Cotte, J.-F., H. Casabianca, S Chardon, J Lheritier et M.-F. Grenier-Loustalot (2004). Chromatographic analysis of sugars applied to the characterisation of mono- floral honey. Analytical and bioanalytical chemistry 380.4, 698–705.
- Cuevas-Glory, L. F., J. A. Pino, L. S. Santiago et E Sauri-Duch (2007). A review of volatile analytical methods for determining the botanical origin of honey. Food Chemistry 103.3, 1032–1043.

- Devillers, J, M Morlot, M. Pham-Delegue et J. Dore (2004). Classification of monofloral honeys based on their quality control data. Food Chemistry 86.2, 305–312.
- Freund, Y., R. Schapire et N Abe (1999). A short introduction to boosting. Journal of Japanese Society For Artificial Intelligence 14.771-780, 1612.
- Kaskoniene, V.,Venskutonis P. R. (2010). Floral markers in honey of various botanical and geographic origins : a review. Comprehensive Reviews in Food Science and Food Safety 9.6, 620–634.
- Nagai, T., R. Inoue, N. Kanamori, N. Suzuki et T. Nagashima (2006). Characterization of honey from different floral sources. Its functional properties and effects of honey species on storage of meat. Food chemistry 97.2, 256–262. R Core Team (2013).
- R : A Language and Environment for Statistical Computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. url : http://www.R-project.org/.
- Zhu, J., H. Zou, S. Rosset et T. Hastie (2009). Multi-class adaboost. Statistics and Its.