grostat 2012

12th European Symposium on Statistical Methods for the Food Industry
Paris, France, (28), 29 February & 1-2 March 2012

CRAGGING: a novel approach for inspecting Italian wine quality

CRAGGING: une nouvelle approche pour évaluer la qualité des vins italiens

Eugenio Brentari
Maurizio Carpita
Marika Vezzoli
Department of Quantitative Methods,
University of Brescia, Italy

Agenda

- Introduction
- Regression Trees
- CRAGGING
- Final Model
- Case Study

Introduction

Assessing the wine quality is a challenging task due to the multifaceted nature of this concept

Subjective evaluations and **objective features** are mixed together in order to get effective ranking of wines

It is important **to identify the fundamental attributes** since they can lead to significant improvements in the production process

Introduction (cont'd)

Using a data mining technique, we inspect the quality of Italian red and white wines →
which variables have a major impact on it?

Since the data have a hierarchical structure (wines grouped with respect to the grapes) we use the **CRAGGING**

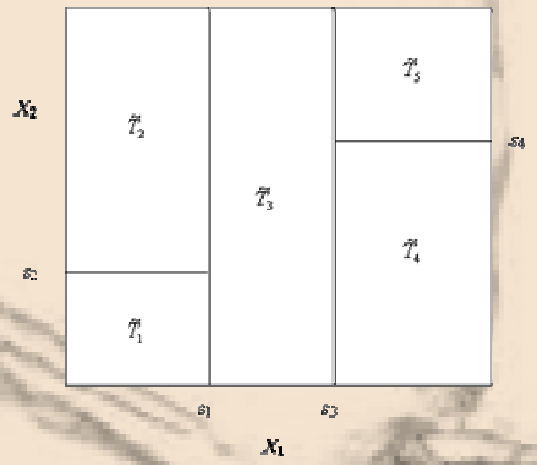
Extracting a simple model from the CRAGGING, we identify the **"true path" towards the quality**

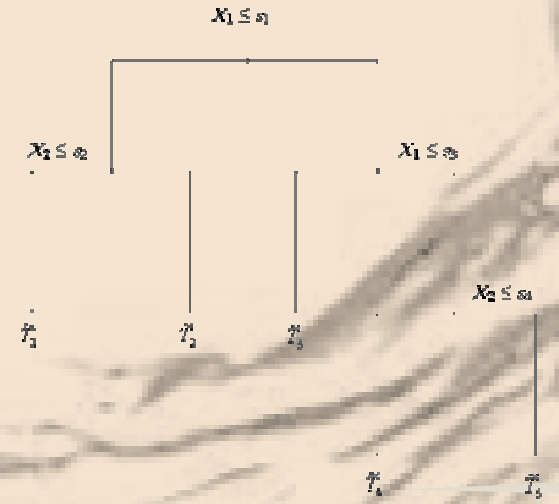
Regression Trees

One of the most famous data mining techniques

Regression Trees

They partition the predictor space into homogeneous subsets with respect to the dependent variable Y





Regression Trees (cont'd)

Advantages

- Fast algorithm
- It can deal with every type of variable
- It provides good results also with → missing values
  → correlated variables
- Interpretability (if trees are small)

Disadvantages

- Non accurate predictor
- **Instability** → small changes in the data determines big changes in the results

CRAGGING

- Combining multiple versions of unstable classifiers (e.g. trees) increases the accuracy of the predictors
- **P&C techniques** (Perturbation & Combination): perturb the training set to generate multiple predictors and combine these by averaging

  Bagging (Breiman, 1996)

  Random Forests (Breiman, 2001)

  Boosting (Freund and Schapire, 1996)

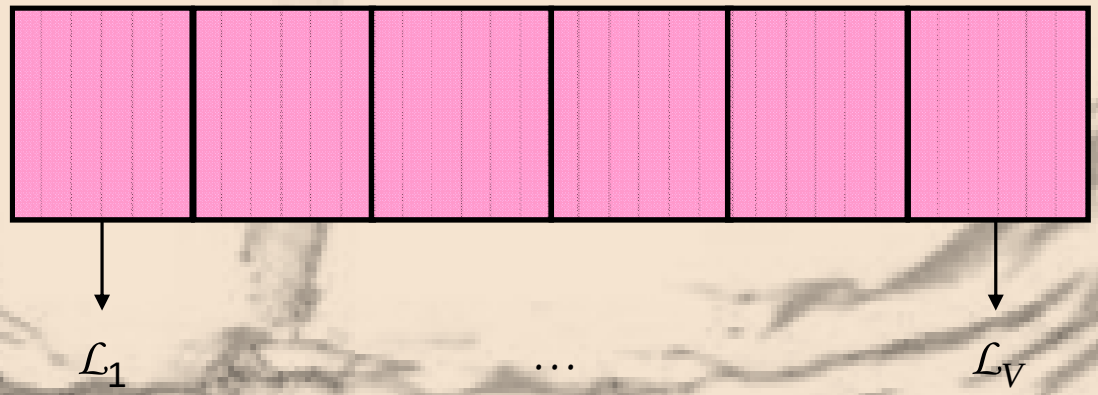
CRAGGING

CRoss-validation **AGG**regat**ING**

(Vezzoli and Stone, 2007)

CRAGGING (cont'd)

Let $(Y,\mathbf{X})$ a database with N observations where Y is the response variable and $\mathbf{X}$ is the matrix of R predictors. The observations are divided in J groups each one composed by n_j observations. Denote with $\mathcal{L}=\{1,2,\ldots,J\}$ the set of groups and with $\mathrm{x}_{ji}=(x_{1ji},x_{2ji},\ldots,x_{rji},\ldots x_{Rji})$ the vector of predictors for i-th subject of group j. The set $\mathcal{L}$ is randomly partitioned in V subsets denoted by $\mathcal{L}_v$, $v = 1,\ldots,V$ each one containing J_v groups.



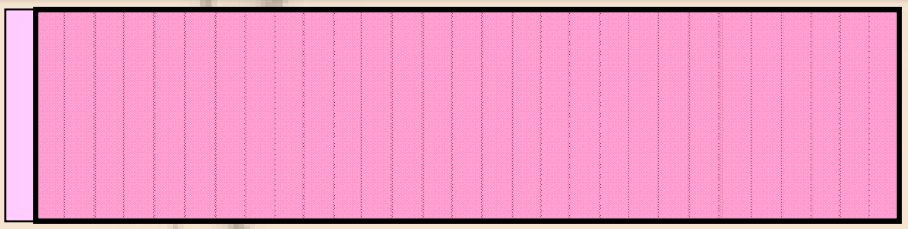
CRAGGING (cont'd)

Training set denoted by $\mathcal{L}_v^c$

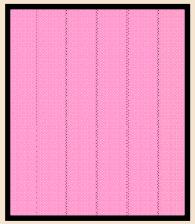
containing J_v^c groups

Wines
Alcamo DOC
Alcamo DOC
Alcamo DOC
...
Gavi DOCG
Gavi DOCG
Gavi DOCG

CRAGGING (cont'd)



Perturbation of the training set



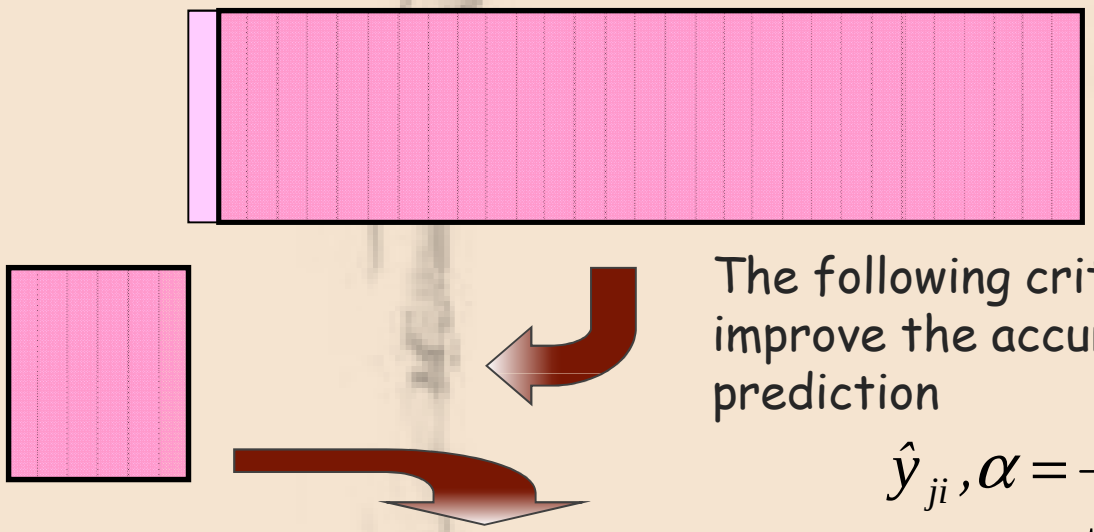
The corresponding prediction in the test set $\mathcal{L}_v$ is

$$\hat{f}_{\alpha,\mathcal{L}^c_{v\setminus l}}(\cdot)$$

Cost-complexity parameter

Wines	$\hat{y}_{ji,\alpha l}$
Alcamo DOC	50.2
Alcamo DOC	40.3
Alcamo DOC	25.6
...	...
Gavi DOCG	60.4
Gavi DOCG	76.2
Gavi DOCG	51.6

CRAGGING (cont'd)



The following criterion is used to improve the accuracy of the prediction

$$\hat{y}_{ji},\alpha = \frac{1}{J_v^c} \sum_{l \in \mathcal{L}_v^c} \hat{y}_{ji,\alpha l}$$

Wines	$\hat{y}_{ji,\alpha l}$	$\hat{y}_{ji,\alpha l}$	...	$\hat{y}_{ji,\alpha}$
Alcamo DOC	50.2	60.5		67.9
Alcamo DOC	40.3	56.8		54.3
Alcamo DOC	25.6	45.7		39.7
...	...	...	...	...
Gavi DOCG	60.4	53.7		65.8
Gavi DOCG	76.2	65.2		64.8
Gavi DOCG	51.6	49.5		54.9

CRAGGING (cont'd)

The procedure is repeated for different values of α and the algorithm chooses the optimal tuning parameter α^*

$$\alpha^\star = \arg\min_{\alpha} L\left(y_{ji}, \hat{y}_{ji,\alpha}\right) \quad \text{with} \quad j \in \mathcal{L}, \quad i = 1,2,...,n_j$$

where $L()$ is a generic loss function. The entire procedure is run M times in order to minimize the generalization error, then averaging the results in order to get the **CRAGGING predictions**:

$$\tilde{y}_{ji}^{\text{crag}} = M^{-1} \sum_{m=1}^{M} \hat{y}_{ji,\alpha^*} \quad \text{with} \quad j \in \mathcal{L}, \quad i = 1,2,...,n_j$$

Final Model

We fit a **Final model** using the average of predictions obtained in the first step in place of the original dependent variable.

The substitution of y with $\hat{y}$ mitigates the effects of noisy data on the estimation process that affect both the predictors and the dependent variable itself. In detail

- The results of the CRAGGING are combined with a single tree:
  - The dependent variable Y is replaced with $\tilde{y}_{ji}^{\text{crag}}$
  - A single tree is grown on 90% of the observations and tested on the remaining 10%. The α^* is used like cost complexity parameter

Case study: Altroconsumo Wines

The analysis was carried on the dataset that **Altroconsumo**, an Italian Independent Consumers' Association, uses for its guide (***Guida Vini 2011***). Each year, Altroconsumo chooses some red and white wines and evaluates their characteristics (in almost all the cases they **cost less than 15 €**).

We analyzed **231 red and white wines** grouped in **49 clusters** defined by the type of grapes used by producers. We focused our attention on **chemical** and **sensory** characteristics, since we expect these features could be the **major drivers of the wine quality**.

We use the chemical and sensory variables as covariates and a score of wine quality (attributed by Altroconsumo) as response variable

Case study: Altroconsumo Wines (cont'd)

Chemical variables

- Alcoholic strength (*Alcohol*)
- Residual sugar (*Residual Sugar*)
- Reducer sugar (*Reducer Sugar*)
- Total acidity (*Acidity tot*)
- Volatile acidity (*Acidity vol*)
- Free/Total sulphur anhydrides (SO_2)
- Total sulphur anhydrides (SO_2 *tot*)

Case study: Altroconsumo Wines (cont'd)

Sensory variables

Experienced judges express their vote about the sensory variables (divided in 4 groups). The perception of each descriptor has been registered using a **0-9 scale**

- **Visual characteristics**
  Color saturation, Green reflection, Gold reflection, Violet reflection, Garnet reflection, Visual sparklingness, Attraency
- **Olfactory characteristics**
  Floral, Fruit, Vegetal, Spicy, Olfactory intensity, Olfactory quality, Olfactory frankness, Perception, Harmony
- **Gustatory characteristics**
  Structure, Harmony, Acidity, Bitterness, Sweet, Astringency, Aromatic Richness
- **Intense Aromatic Persistence**
  Persistence, Frankness, Quality

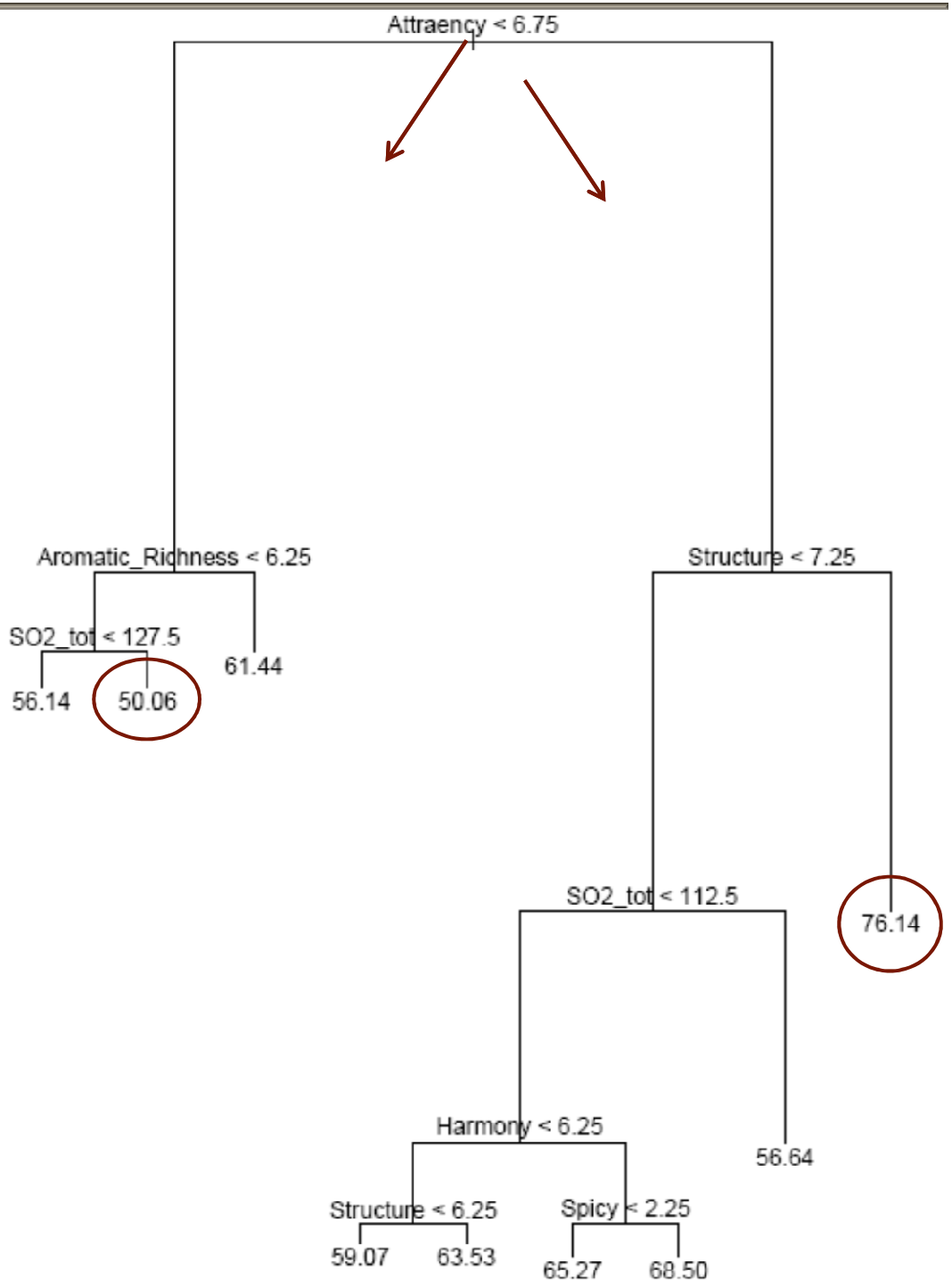
$Y \rightarrow$ Composite score: 0 (lowest quality) 100 (highest quality)

Result

The Final Model could be used as a tool to drive the quality towards high scores

Wines with highest score have a **pleasant aspect** and a **good flavour**

Wines with lowest score **do not have an attractive aspect, do not have a pleasant flavour,** and the **level of the total sulphur anhydrides is very high**



Main References

Brentari, E., Levaggi, R. (2010). Hedonic price for Italian Red Wine: a panel analysis. *Proceedings of the 11th European Symposium on Statistical Methods for the Food Industry*, Academy School, Afragola (Napoli), pp. 249-258.

Brentari, E., Levaggi, R., Zuccolotto, P. (2011). Pricing strategies for Italian red wine. *Food Quality and Preference*, 22, pp. 725-732.

Brentari, E., Zuccolotto, P. (2010). The implicit value of chemical and sensorial quality in the hedonic analysis of low-priced Italian red wines. *Proceedings of 11th European Symposium on Statistical Methods for the Food Industry*, Academy School, Afragola (Napoli), pp. 269-276.

Vezzoli, M., Stone, C. J. (2007). CRAGGING. In Book of Short Papers CLADAG (Classification and Data Analysis Group) 2007, pp. 363-366, EUM.

Vezzoli, M. (2011). Exploring the facets of overall job satisfaction through a novel ensemble learning. Electronic Journal of Applied Statistical Analysis, 4(1), pp. 23-38.

Vezzoli, M., Zuccolotto, P. (2011). CRAGGING measures of variable importance for data with hierarchical structure. In New Perspectives in Statistical Modeling and Data Analysis, pp. 393-400, S. Ingrassia, R. Rocci, M. Vichi (Eds.), Springer.


brentari@eco.unibs.it
carpita@eco.unibs.it
vezzoli@eco.unibs.it
Cheers!