Applying Machine Learning to High-Quality Wine Identification

Giorgio Leonardi and Luigi Portinale, *Computer Science Institute, DISIT*

UNIVERSITÀ DEL PIEMONTE ORIENTALE

Problem setting

The project ***TRAQUASwine*** aims at defining methods for data analytical assessment of the authenticity and protection against fake versions of some of the highest value **Nebbiolo-based** wines from Piedmont region in Italy.

Wine adulteration leads to:

- Health problems
- Big economic damage (several millions of Euros each year)

"Classical" fraud detection:

- Complex chemical profiles (e.g. "terroir" influence)
- Very expensive instruments and analyses
- Very large databases needed

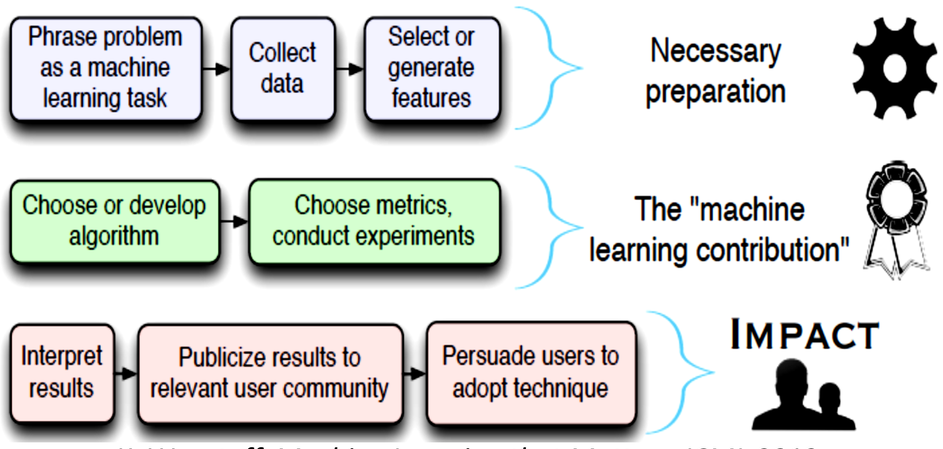
Objectives of this work

Show that the problem can be addressed without expensive and hyper-specialized wine analyses

Demonstrate usefulness of data mining classification algorithms on the resulting chemical profiles

METHODS AND DATA

The proposed approach can be regarded as an instance of the 3-phases Wagstaff's scheme



K. Wagstaff. Machine Learning that Matters, ICML 2012

Data Preparation

- Different wines from different areas and different wineries
- Different typology of wine (commercial vs model-wines)
- Use of common chemical parameters (cheap analyses): no need for pluriennal consolidated datases
  - *spectrophotometric and chromatographic methods*

Wine	#Samples	Category	Wine	# Samples	Category
Barolo (BRL)	18	HQ	Gattinara (GAT)	6	HQ
Barbaresco (BRB)	18	HQ	Ghemme (GHE)	26 (8)	
Nebbiolo (NEB)	34 (16)		Sizzano (SIZ)	6	
Roero (ROE)	6		GAT Blend (BLE)	10	CTRL
Langhe (LAN)	22 (10)	CTRL	No NEB (NON)	12	2 LEV CTRL

146 samples for 9 classes; 12 samples for 2nd level control (no training)

UNSUPERVISED ANALYSIS

Data: 40 continuous features (chemical parameters)
missing values (in about 1/3 of the samples)
9 wine types

Problem: Clustering

- multi-class evaluation: K-means with K = 9 clusters (wine types), EM with no predefined no. of clusters, EM with predefined set of 9 clusters;
- binary evaluation: K-means and EM with predefined set of K = 2 clusters

Datasets: Original: 146 samples, 40 continuous features
PCA: 146 samples; dimensionality reduction through PCA

	K-means				EM					
	Original		*PCA*		*Original*			*PCA*		
	nc=9	nc=2	nc=9	nc=2	nc=6	nc=9	nc=2	nc=13	nc=9	nc=2
Purity	0.43	0.71	0.42	0.71	0.41	0.53	0.71	0.55	0.51	0.71
Precision	0.21	0.59	0.24	0.58	0.22	0.33	0.59	0.37	0.39	0.58
Recall	0.41	0.52	0.33	0.78	0.49	0.37	0.52	0.24	0.37	0.52
Rand Index	0.69	0.50	0.75	0.53	0.68	0.80	0.50	0.83	0.83	0.50
F1-measure	0.28	0.55	0.28	0.66	0.30	0.34	0.55	0.29	0.38	0.55
FM Index	0.29	0.55	0.28	0.67	0.33	0.35	0.55	0.30	0.38	0.55

- Cluster quality indices generally poor. No significant difference between original and PCA dataset
- No predefined no. of clusters: EM produces nc=6 clusters on the original dataset and nc=13 clusters on the PCA dataset (out of 9 actual classes).
- These experiments suggested that regularities associated to wine classes should be better investigated through supervised methods.

SUPERVISED ANALYSIS

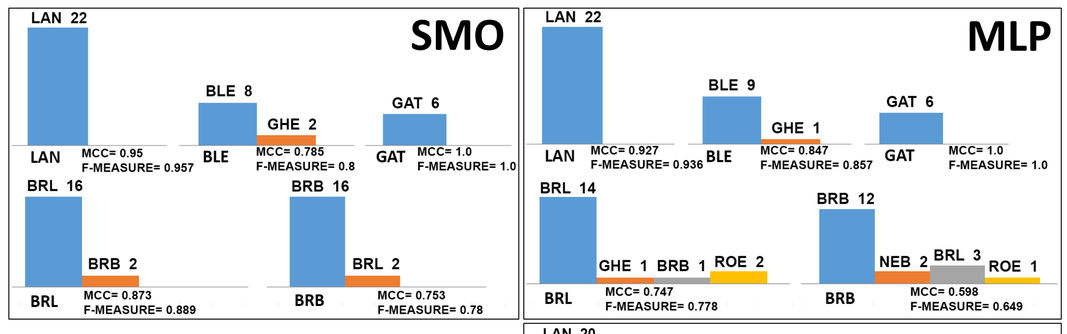
Data: 13 continuous features (out of the original 40); 9 wine types

Problem: Classification (9 classes);
Emphasis on misclassification of HQ classes wrt CTRL

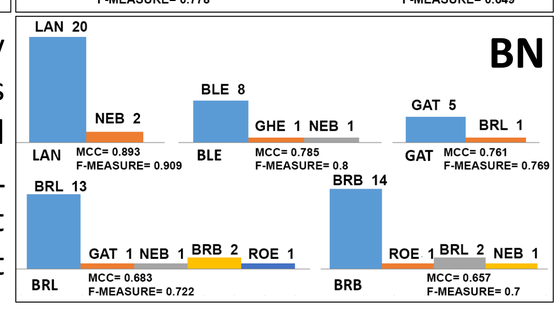
- Bayes Net Classifier (**BN**): standard Cooper/Herskovits algorithm; max. 3 parents per node
- SVM based on **SMO** (Sequential Minimal Optimization) and Pearson Universal Kernel - Platt's scaling for output probabilities
- Multi Layer Perceptron (**MLP**) with one hidden layer
  Hidden units = (#features + #classes) / 2

Datasets: D1: 146 samples, 13 features
D2: 14600 artificial samples, generated from trained BN model

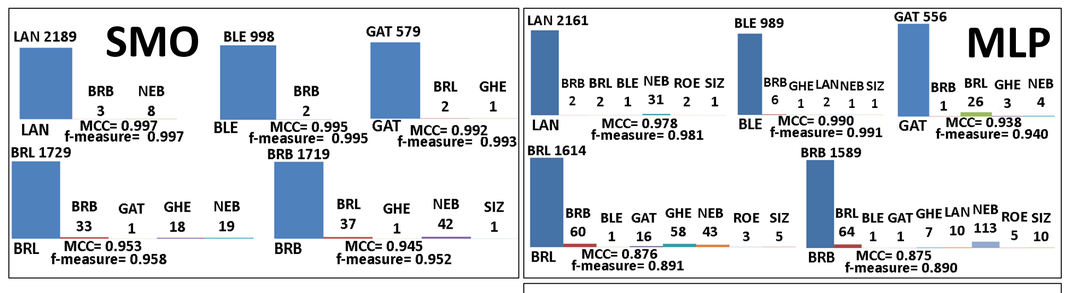
Dataset D1



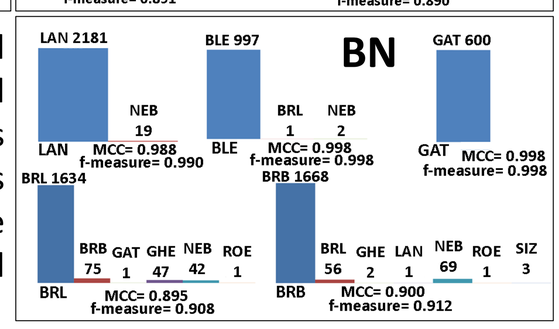
Control wines (LAN, BLE) classified with very good accuracy. No CTRL wine misclassified as HQ. HQ wines (BRL, BRB, GAT) also recognized very well, with no misclassification of CRTL wines. The few misclassifications are not surprising, pointing to very similar wines wrt origin, production and grape composition.



Dataset D2



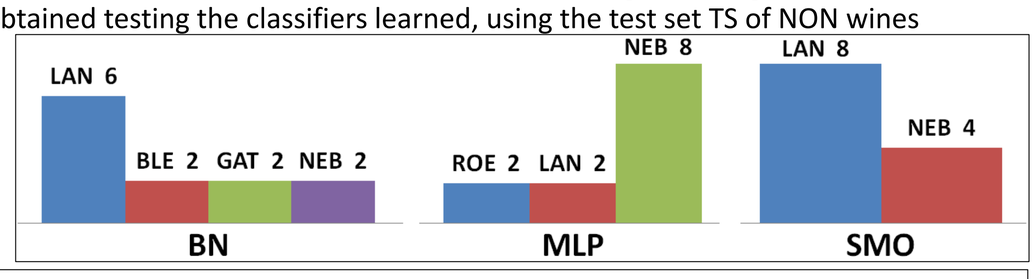
As in D1, no CTRL wine predicted as HQ and vice versa. SMO and BN have the same good performance as in D1. MLP slightly augments the no. of misclassifications, but still performs good. Also for the synthetic dataset, the tested classifiers represent a valuable and viable tool for the goal of the study.



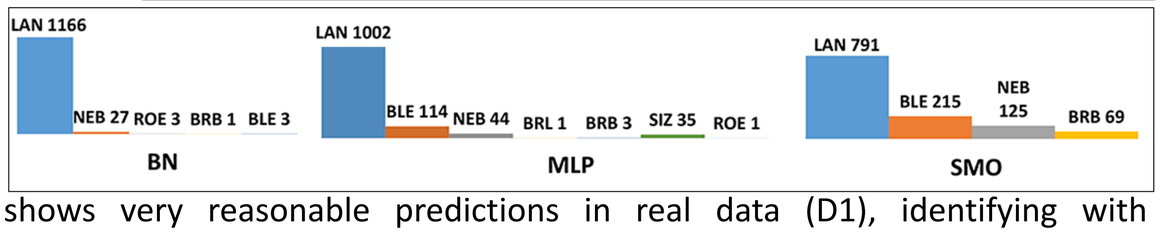
Prediction of non-wines (NON)

Results obtained testing the classifiers learned, using the test set TS of NON wines

D1



D2



SMO shows very reasonable predictions in real data (D1), identifying with acceptable performance the only class (LAN) containing percentages of the grapes present in the test cases. Looking at D2, instead, BN is the more robust classifier. In both cases, supervised classification techniques are very promising in avoiding incorrect introductions of unrelated cultivars.

CONCLUSION. The results of the experiments suggest that standard chemical profiling of Piedmont Nebbiolo-based wines, coupled with data mining classification techniques, can be a powerful tool to authenticate high-quality and high-value wines. All the tested classifiers performed rather well with respect to the objectives of the work, with BN being the more problematic in some situations, and MLP comparable in performance with SMO, the latter showing a better robustness with respect to possible fake wines.