A HYBRID WINE CLASSIFICATION MODEL FOR QUALITY PREDICTION

Terry Hui-Ye Chiu,
Chien-Wen Wu,
Chun-Hao Chen

Introduction

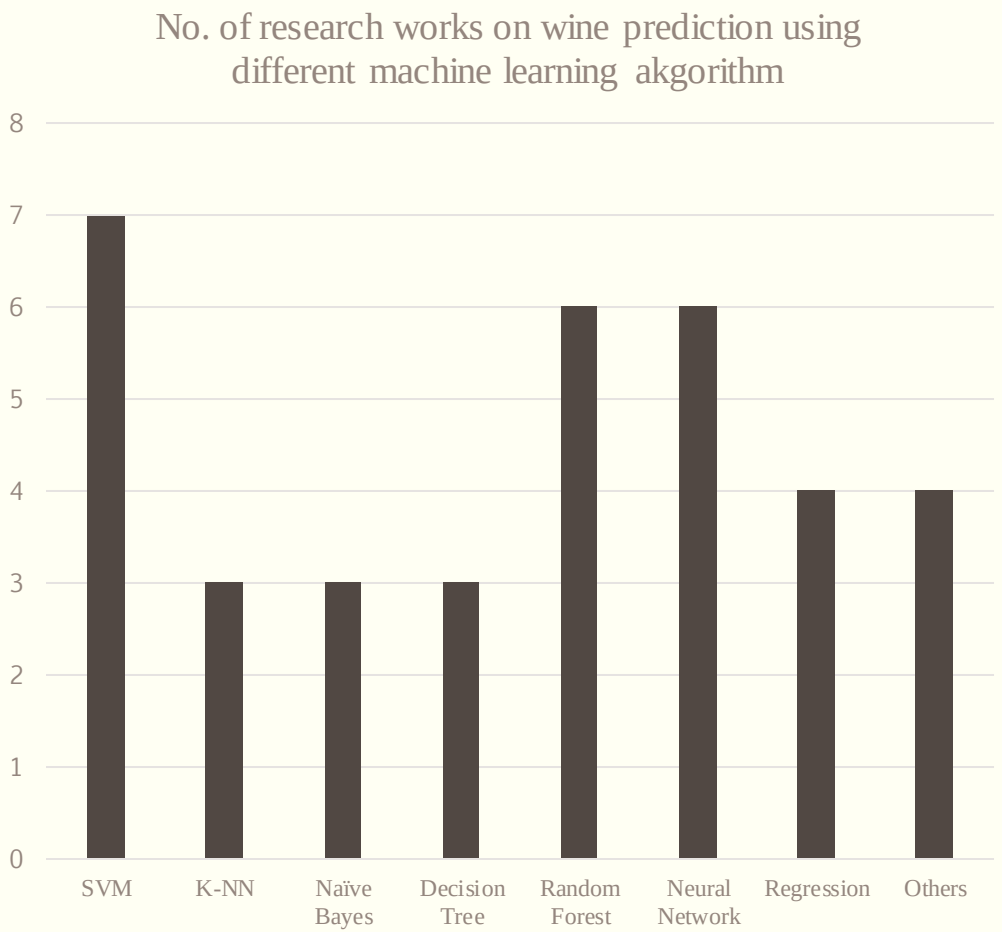
- Wine is an exciting and complex product
  - Distinctive qualities
- Wine testing is complex and diverse.
  - Currently based on opinion of wine expert
  - The opinion of a wine expert is influential
    - Costly and subjective
- With introduction of new technology
  - Machine learning techniques
    - Assess the quality of wine
    - Determine what attributes make a "good" wine that the satisfy consumers
  - Mostly focused on analyzing different classifiers to find the “best” classifier
  - Restrained to certain type of dataset

Introduction cont.

- Past works mostly focused on
  - Using or comparing different machine learning models
    - Finding which what is best prediction result for specific datasets
- Our goal
- Find a more effective classifier
  - Proposed a hybrid model
    - Consists at least two classifiers at least
    - Random forest, support vector machine etc
- To evaluate the performance of the proposed hybrid model
  - Experiments also made on the wine dataset (Cortez et al., 2009)

Background Research

- Based on research during 2009 – 2020
- Most popular
  - SVM, Random Forest and Neural Network
- Classifier used
  - SVM and Random Forest
  - Why not Neural Network?
    - Too many types of NN



Support vector machine (SVM)

- Supervised machine learning model for solving a classification problem
- The central concept of SVM is
  - Utilized the kernel function to find the hyperplane that can separate instances into categories
- Three hyperparameters in SVM that include
  - Penalty factor C
    - Regularization parameter
    - Controls the trade-off between maximizing the margin and minimizing the training error
  - Parameter gamma γ
    - Defines how far the influence of a single training example reaches
  - Kernel function
    - Three different main types of kernels:
      - Linear
      - Poly
      - Rbf

Random Forest

- Ensemble of individual decision trees
- Each individual tree in the random forest given out a prediction
  - Class with most votes becomes final prediction result
- Remove the short-coming of decision tree
- The random forest has six hyperparameters:
  - No. of estimators
    - Number of trees to create
  - Maximum features
    - Amount of features need to be selected for determining the split
  - Maximum depth
    - How many level the tree has to be expanded down to each node till reach leaf node
  - Minimum samples split
    - Minimum number of elements in each node required to stop further splitting
  - Minimum samples leaf
    - Minimum number of instances allowed in a leaf node
  - Bootstrap
    - Sampling instances with or without replacement.

Proposed Hybrid Wine Classification Model

Hyperparameters are fitted to the model, and the model is then evaluated using the predefined criteria (mainly accuracy).

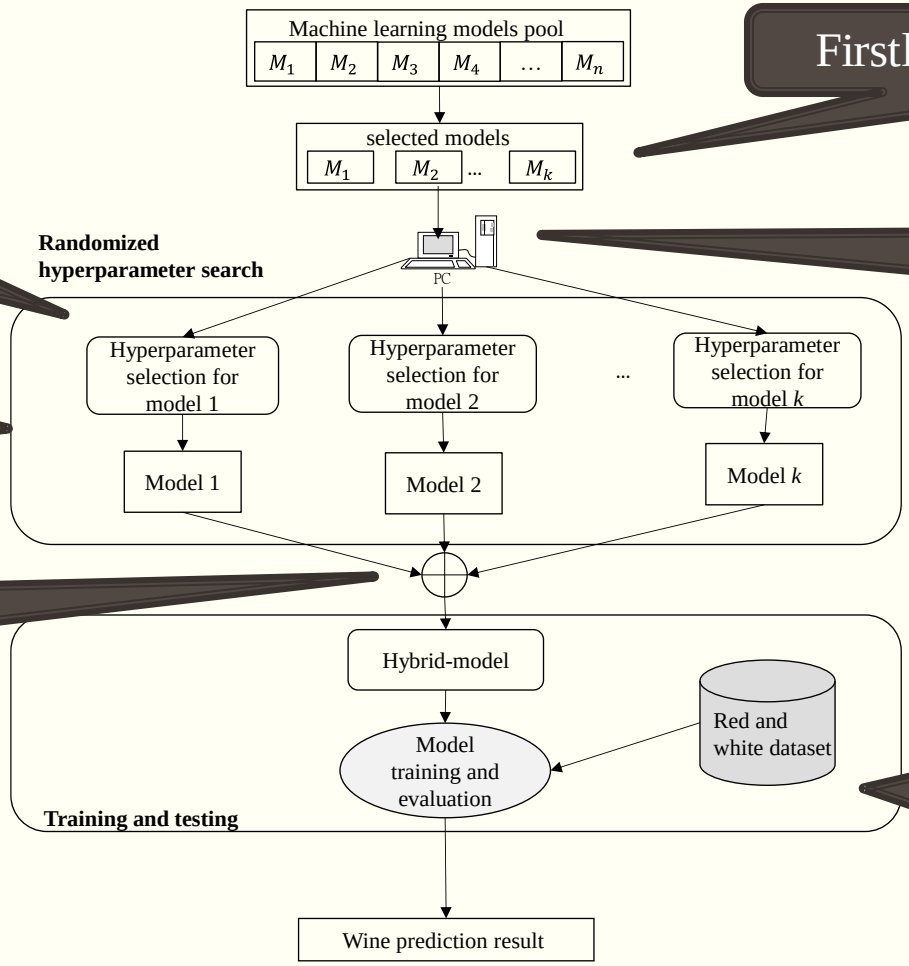
Firstly selected at least two models

Initial ranges of the hyperparameters associated with each model are set

The modification process of trial and error to find the best setting based on performance.

After the tuning procedure, the selected models are merged to form a hybrid model

The model is then trained and tested for *n* times with different training and testing data for each iteration



Experimental Evaluation – Data Description

- The wine dataset from the UCI database (Cortez et al., 2009)
- Two sets of wine data (red and white)
  - Red wine contains 1599 instances,
  - White wine contains 4898 instances.
- Both datasets contain 11 physiochemical variables
  - Including fixed acidity, volatile acidity, citric acid, residual sugar,
  - chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, Sulphates, and alcohol.
  - The output data (sensory data) is a quality rating from 0 (very bad) to 10 (excellent).

Performance Measure Metrics

- The four criteria:
- Accuracy
  - $accuracy = \frac{TP+TN}{TP+TN+FP+FN'}$
- In the multi-class classification problem:
  - Macro-averaging measurements for precision, recall and f1 can provide a clearer reference
- Precision
  - $\text{Macro} - \text{Precision}\ MP_{Classifier} = \frac{\sum_c P_c}{number\ of\ classes}.$
- Recall
  - $\text{Macro} - \text{Recall}\ \text{MR}_{classifer} = \frac{\sum_c R_c}{number\ of\ classes}$
- F1-score
  - $\text{Macroi} - \text{F1} \quad \text{F1}_{classifer} = \frac{\sum_c \text{F1}_c}{number\ of\ classes}$

Experimental Analysis

- Since most of the past works mainly focus on accuracy
  - Only compared the accuracy of the proposed model against others
- Also both works set the training and testing datasets ratio to 80/20
  - The proposed model also use 80/20 for training and testing
- Proposed method accuracy on both wine set are better than both models

	Accuracy	
Models	Red Wine	White Wine
Cortez et. al [6]	0.45	0.51
Apalasamy et. al [17]	0.62	0.65
Proposed model	0.71	0.68

Experimental Analysis – Change of ratio

- To examine further on the performance of the proposed model
  - Exam model under different training and testing data ratio
  - Under different evaluation indicator

	Red wine			White wine		
Testing dataset size percentage	10%	20%	30%	10%	20%	30%
Accuracy	0.69	0.71	0.66	0.70	0.68	0.67
Macro-Precision	0.42	0.43	0.34	0.42	0.37	0.47
Macro-Recall	0.41	0.38	0.32	0.41	0.34	0.35
Macro-F1 score	0.41	0.39	0.32	0.41	0.34	0.36

Conclusion and Future Work

- This paper has proposed a hybrid wine classification model for quality prediction
- The proposed algorithm first selects *n* models from the given model pool.
  - Then, the hyperparameters are then searched by the randomized search method.
  - The models with acceptable performances are merged as the hybrid model.
- In the future, we will continue to design an more robust algorithm
  - Hybrid models
  - Hyperparameters for any wine dataset tuned by using evolutionary algorithms.