ISSN 2616-5775 Vol. 7, Issue 6: 70-75, DOI: 10.25236/AJCIS.2024.070611

Research on Wine Quality Prediction with HHO-SVM

Xinyang Li

College of Electronics and Information Engineering, Sichuan University, Chengdu, 610065, China

Abstract: *Quality certification of wines of different qualities is important, and advances in quality certification require innovative models to further improve accuracy. This study incorporates a substantial dataset comprising 4898 samples and 12 physicochemical variables. Employing MIN-MAX normalization as a foundation, the primary focus is on utilizing the Support Vector Machine (SVM) optimized by Harris Hawks Optimization (HHO). HHO optimized the hyperparameters of the SVM to achieve a nearly thirty percent increase in model accuracy, in addition to this, comparative analysis reveals that HHO-SVM outperforms other models, including Decision Tree, AdaBoost, Backpropagation (BP) neural network, Naive Bayes, Logistic Regression, and the conventional SVM. Empirical findings from extensive experimentation on the WINE dataset demonstrate the achievement of a remarkable 100% accuracy by the model. The results shown that the HHO-SVM holds the potential to elevate wine production, thereby positively impacting the wine industry.*

Keywords: *Support Vector Machine, Harris Hawks Optimization, MIN-MAX normalization, Vinho Verde, Wine quality*

1. Introduction

Nowadays, wine has been socially valued in many senses and been viewed as a necessity in human interaction as well as a luxury good to some extent. Vinho Verde wine originates in northwestern Portugal, bordering the Atlantic Ocean, exports of which have been increased by 36% from 1997 to 2007[1]. To further sustain and increase the benefits of this growth, wine industries is dedicated in studying and applying new technologies, making wine certification and quality assessment inevitable. Certification plays a vital role in preventing the illegal adulteration and guaranteeing the quality, in the meantime, as a part of the certification, quality assessment is conducive to the improvement of wine making, which is an indispensable element in ensuring the overall integrity and excellence of the certified products.

The advent of machine learning makes it possible to create a model from data so as to predict wine quality in a better way. In 1991, a "Wine" dataset which contains 178 instances with measurements of 13 chemical constitutes was donated into UCI repository and was used to distinguish among three cultivars originating from Italy. The dataset is primarily utilized as a benchmark for new DM classifiers owing to its easy discrimination. Piyush et al.[2] adopted AdaBoost, XGBoost and random forest (RF) to predict New Zealand Pinot noir wines. Recently in 2016,Yesim et al[3]. used KNN, GF and SVM to predict wine quality. In 2017, Yogesh Gupta[4] used NN and SVM to predict wine quality with different 11 physicochemical characteristics. Moreover, Paladugu Sirivanth[5] used a SVM algorithm to assess the excellence of wine by selecting crucial aspects that is instrumental in defining the quality of the wine. Dipak Kumar Jana[6] utilized 178 samples with 13 different physicochemical characters, five neural networks and six support vector methods for wine quality certification, among which Quadratic support vector machine outperformed.

SVM is a popular supervised learning algorithm utilized in machine learning for both classification and regression tasks, when combined with intelligent optimization algorithm[7] the performance and generalization capability of SVM are improved. In 2011, SSA-SVM was adopted to forecast short-term rainfall . In the year 2021, Huang and Jiang[8] utilized AO-SVM ,LSTM, GRNN, SSA-SVM, OCSSA-SVM, PSO-SVM and WOA-SVM to predict the soil moisture content. The next year, Zhou et al.[9] used WOA-SVM combined with a portable electronic nose system to predicate tomato storage quality. Chen et al.[10] realized the early fault diagnosis of bearing of a vibrating screen exciter bearing by aquila optimizer improved SVM (AO-SVM).

ISSN 2616-5775 Vol. 7, Issue 6: 70-75, DOI: 10.25236/AJCIS.2024.070611

There is still much room for improvement, even if some academics have assessed wine quality above using machine learning methods. In this work, the hyperparameters of SVM are optimized by HHO to enhance the precision of the model, and the combined application of SVM and HHO is studied in depth and compared with previous quality prediction and machine learning methods.

This paper utilized MIN-MAX standardization to preprocessing the data, and HHO-SVM are executed to predict the wine quality with different 11 physicochemical characteristics. The paper is recognized as follows; Section 2 presents the wine data, MIN-MAX normalization, SVM models and HHO; in Section 3, the experiment design is outlined and the obtained results are analyzed and compared with previous researches; in Section 4, final conclusions are presented.

2. Materials and methods

2.1 Wine data

The dataset can be downloaded from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/109/wine) and is associated with white variant of the Portuguese "Vinho Verde" wine, which was gathered from May/2004 to February/2007, exclusively from protected designation of origin samples that underwent tasting at the official certification entity (CVRVV). The dataset is composed of 4898 samples and each sample consists of 12 physicochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality rating, among which the quality rating is on the basis of a sensory evaluation conducted by a minimum of three sommeliers, graded across 11 quality levels ranging from 0-very poor to 10-very excellent, typically falling within the range of four to seven.

The goal of the dataset is to predict the quality of the "Vinho Verde" wine, using a range of physicochemical. In order to facilitate parameter tuning and model selection, we divide eighty percent of the dataset into a training set and the remaining twenty percent into a test set.

2.2 MIN-MAX Normalization

To avoid large level gaps between evaluation indicators that would reduce the accuracy of experimental results, normalization is often utilized in experiments related to data mining. Normalization refers to scaling data so that it falls into a small specific interval, which aims to eliminate unit restrictions from data and converting them to pure dimensionless values facilitate the ability to compare and weight indicators of different units or magnitudes.

MIN-MAX normalization is a way of data standardization[11], and it is also known as deviation normalization, involves linearly transforming the initial data to fit within the interval [0,1]. Apply this rransformation to the sequence x:

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where max represents the highest value within the sample data, and min represents the lower value within sample data. The transformation yields a new sequence $y \in [0,1]$ with no measure.

2.3 Support Vector Machine

SVM is a binary classification model derived from statistical learning theory by Vapnik and Chervonenkis and first introduced in 1992 by Boser et al. of which fundamental is a linear classifier that maximizes the intervals within the feature space, and the interval maximization distinguishes it from the perceptual machine.

The fundamental concept behind the SVM method is to determine a separating hyperplane that effectively divides the training dataset while maximizing the geometric margin.

2.4 Harris Hawks Optimization

Harris Hawks adopt a "seven kills" strategy, in which several hawks attempt to collaboratively attack a detected rabbit that trying to escape outside the river, approaching from various directions simultaneously. Harris Hawks Optimization (HHO), a novel nature-inspired optimization paradigm, is inspired by the hunting behaviors of Harris hawks, including exploration of a prey, surprise pouncing,

ISSN 2616-5775 Vol. 7, Issue 6: 70-75, DOI: 10.25236/AJCIS.2024.070611

and utilization of various attacking strategies. The implementation of the HHO algorithm can be roughly categorized into three stages and can be described below.

2.4.1 Exploration phase

The hawks spend hours waiting, observing, and monitoring the desert site to detect a prey which cannot be seen easily. In HHO, Harris hawks are the candidates for scenarios, and the best candidate for each process is treated as either the expected prey or something similar to it. Harris hawks select random perching spots and two strategies to detect potential prey. If the chances q for each perching strategy are equal, the Harris hawk will perch according to the location of other members and prey when $q < 0.5$; when $q > 0.5$, the Harris hawks will randomly perch on a large tree within the range of the flock's activity, as modeled by.

2.4.2 Transition from exploration to exploitation

The HHO algorithm transitions from exploration to exploitation, and then converted between different exploitation behaviors based on the prey's escape energy. During the escape, the energy will greatly reduce.

2.4.3 Exploration phase

This phase can be divided into four strategies, which can be concluded as soft besiege, hard besiege, soft besiege with progressive rapid dives and hard besiege with progressive rapid dives. The choice of strategy is based on the value of r and the absolute value of E, where r is a chance for the prey to escape before the surprise attack and E is the escape energy of the prey.

2.5 Model evaluation

To assess the performance of the trained models, four evaluation metrics were employed: accuracy, precision, recall, and F1 score. These metrics evaluate various aspects of the model's performance. A true positive (TP) indicates correct identification of a positive label, while a false negative (FN) represents a missed identification. Conversely, a true negative (TN) indicates correct identification of a negative label, and a false positive (FP) signifies the incorrect identification of a positive label. By considering these metrics, a thorough assessment of the models' performance can be obtained.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Presion = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}Score = \frac{2Presion \bullet Recall}{Presion + Recall} \tag{5}$$

3. Results and discussions

3.1 Comparison of models

Other data mining methods such as Decision Tree, AdaBoost, BP neural network, Naive Bayes and Logistic Regression are adopted to predict the wine quality to make comparison. The models are introduced as follows:

Decision Tree: The decision tree, composed of root, internal and leaf nodes is constructed to assess the likelihood that the anticipated value of the net present value (NPV) exceeds or equals zero. This evaluation is based on the known probability of different scenarios occurring. The generation of the decision tree can be primarily split into two parts: nodes splitting and thresholds determination.

AdaBoost: AdaBoost is an iterative algorithm designed to train various classifiers, with a focus on weak classifiers, using the same training set,,It then ensembles these weak classifiers to construct a

ISSN 2616-5775 Vol. 7, Issue 6: 70-75, DOI: 10.25236/AJCIS.2024.070611

more robust final classifier. The algorithm achieves this by adjusting the weights assigned to each sample based on the correctness of the previous classifiers. The revised data, featuring adjusted weights, is forwarded to the lower level classifiers for training. Subsequently, the classifiers derived from each training are merged to form the ultimate decision classifier.

BP neural network: BP neural network, proposed by a team of scientists led by Rumelhart and McCelland, is a multilayer feed forward network trained using the error backpropagation algorithm. This network possesses the ability to learn and store a considerable number of input-output pattern mapping relationships.

Naive Bayes: Naive Bayes is a classification method grounded in Bayes' theorem, assuming that features are independent under the given conditions. It learns the joint probability distribution of inputs and outputs from training data, assuming the features are conditionally independent, and then calculates the maximum a posteriori probability for new instances using Bayes' theorem. Instead of directly learning the joint probability distribution of inputs and outputs, this method accomplishes this by learning the class prior probabilities and class conditional probabilities.

Logistic Regression (LR): Logistic Regression is a machine learning method designed for addressing binary classification problems by estimating the likelihood of an event. In logistic regression, the dependent variable y adheres to a Bernoulli distribution. While linear regression is based on the assumption that the dependent variable y follows a Gaussian distribution, so logistic regression is theoretically sound in the context of linear regression. However logistic regression introduces a nonlinear element through the Sigmoid function, enabling effective handling of 0/1 classification problems.

3.2 Analysis of results

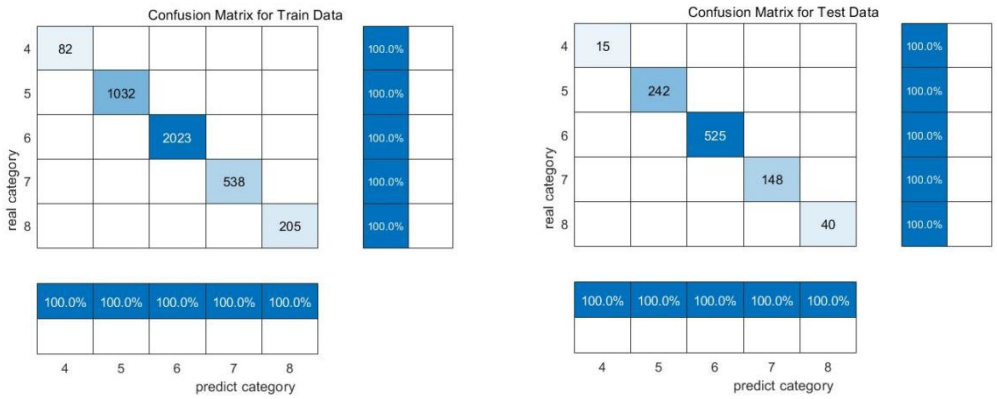

Figure 1: Confusion Matrix for wine data

The confusion matrix, also referred to as the error matrix, is a standard representation for assessing accuracy, displayed in an n-by-n matrix. Each column of the confusion matrix corresponds to a predicted category, showing the total number of data predicted to be in that category. Each row signifies the true attributed classification of the data, indicating the entirety of data instances in that classification.

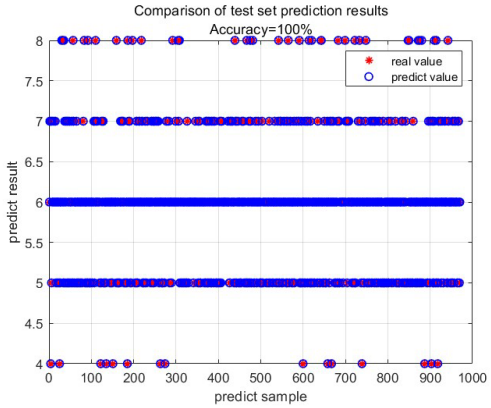

Figure 2: Comparison of test set prediction results

ISSN 2616-5775 Vol. 7, Issue 6: 70-75, DOI: 10.25236/AJCIS.2024.070611

From the confusion matrix (Fig. 1) with Fig. 2 above can be seen that the accuracy of predicting the quality of Vinho Verde using this model is 100 percent. Table 1 presents the results of predicting the quality of the Vinho Verde using different classifiers before data processing.

Table 1: Comparison of different models result (before preparation)

Models	Accuracy	Precision	Recall	F1
Decision Tree	0.45	0.45	0.5182	0.4725
AdaBoost	0.49	0.49	0.4164	0.4332
BP neural network	0.7133	0.7133	0.7078	0.6976
Naive Bayes	0.5912	0.5912	0.5928	0.5801
LR	0.717	0.717	0.7135	0.702
HHO-SVM	1	1	1	1

The most successful prediction of the quality of Vinho Verde was obtained by HHO-SVM using unprocessed data. The accuracy, precision, recall and F1 of the result are both 1. Table 1 clearly shows that HHO-SVM outperforms from the other algorithms. Table 2 illustrates the results of predicting the quality of the Vinho Verde using different classifiers after data processing.

Table 2: Comparison of different models result (after preparation)

Models	Accuracy	Precision	Recall	F1
Decision Tree	0.973	0.973	0.9819	0.9763
AdaBoost	0.5346	0.5346	0.3145	0.3712
BP neural network	0.687	0.687	0.688	0.6728
Naive Bayes	0.5925	0.5925	0.5944	0.5836
LR	0.7262	0.7262	0.7239	0.7105
HHO-SVM	1	1	1	1

After MIN-MAX normalization, the results of the Decision Tree and AdaBoost were significantly improved. The accuracy of result obtained by Decision Tree was increased by 52.3%, and the accuracy of the result obtained by AdaBoost has increased by 4.46%. The most successful prediction of the quality of Vinho Verde was obtained by HHO-SVM using unprocessed data. The Accuracy, Precision, Recall and F1 of the result are both 1. Then, for increasing of prediction success in this study, the raw data was MIN-MAX normalized and the process wine quality classification was repeated by using Decision Tree, AdaBoost, BP neural network, Logistic Regression and HHO-SVM.

Table 3 shows the evaluating values of SVM model before and after the introduction of the intelligent optimization algorithm, i.e. Harris Hawks Optimization, and compares the results of this paper with Piyush Bhardwaj, and Yesim Er's experiments. Piyush Bhardwaj used the SMOTH method to generate 1381 samples from twelve original samples and utilized essential variables (referred as important attributes) to predict the wine quality. In Yesim Er's study, the use of principal component analysis was observed, which indeed improved accuracy and precision.

Table 3: Comparison of model accuracy under different treatments

Models	Accuracy	Precision	Recall	F1
SVM(normalized)	0.6988	0.6988	0.7414	0.6803
HHO-SVM	1	1	1	1
SVM(Piyush)	0.83	0.83	0.88	0.83
SVM(Yesim)	0.478	0.478	0.569	0.519

As a consequence, it can be claimed that SVM model optimized by Harris Hawks Optimization was used for the best results, which makes it possible to categorize the wine quality with a high degree of Accuracy, Precision, Recall, and F1.

4. Conclusions

In conclusion, the growing emphasis on wine quality in recent years is pivotal for upholding consistent high standards across batches, enhancing enterprise reputation, and expanding market share. Beyond its impact on customer satisfaction and loyalty, effective wine quality prediction and testing also play a vital role in reducing return rates and safeguarding the economic interests of enterprises.

This study focuses on precise wine quality prediction utilizing 12 physicochemical variables, employing a dataset of 4898 Vinho Verde samples from northwest Portugal. The wine quality, graded on

ISSN 2616-5775 Vol. 7, Issue 6: 70-75, DOI: 10.25236/AJCIS.2024.070611

a scale of 0 to 10, predominantly falls within the range of 4 to 7. Evaluation metrics such as correctly classified instances, precision, recall, and F1 are used to present the results. Implementing MIN-MAX normalization significantly improved algorithm performance, with the Decision Tree demonstrating the most substantial enhancement—an impressive 44.097% accuracy development, surging from 45% to 97.3%. Comparing classifiers including Decision Tree, AdaBoost, BP neural network, Naive Bayes, Logistic Regression, and HHO-SVM, the experiments reveal that HHO-SVM excels in classification tasks, outperforming the other five algorithms and normal SVM models with 100 percent accuracy. The significance of this work extends to both wine industries and quality assurance. The data-driven approach, grounded in objective factors, yields more objective and accurate results compared to expert opinions influenced by subjective factors. The key features lie in the meticulous data preprocessing procedure and the efficacy of intelligent optimization algorithms.

Looking ahead, larger datasets will be employed for experiments, and exploration of additional machine learning techniques and intelligent optimization algorithms is anticipated for advancing wine quality prediction.

References

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision support systems, 47(4), 547-553.
[2] Bhardwaj, P., Tiwari, P., Olejar Jr, K., Parr, W., & Kulasiri, D. (2022). A machine learning application in wine quality prediction. Machine Learning with Applications, 8, 100261.
[3] Er, Y., & Atasoy, A. (2016). The classification of white wine and red wine according to their physicochemical qualities. International Journal of Intelligent Systems and Applications in Engineering, 4(Special Issue-1), 23-26.
[4] Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, 305-312.
[5] Sirivanth, P., Rao, N. K., Manduva, J., Sekhar, G. C., Tajeswi, M., Veeresh, C., & Kaushik, J. V. (2021, December). A SVM Based Wine Superiority Estimatation Using Advanced ML Techniques. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 207-211). IEEE.
[6] Jana, D. K., Bhunia, P., Adhikary, S. D., & Mishra, A. (2023). Analyzing of salient features and classification of wine type based on quality through various neural network and support vector machine classifiers. Results in Control and Optimization, 11, 100219.
[7] Lian, J., & Hui, G. (2023). Human Evolutionary Optimization Algorithm. Expert Systems with Applications, 122638.
[8] Huang, Y., & Jiang, H. (2022). Soil Moisture Content Prediction Model for Tea Plantations Based on a Wireless Sensor Network. Journal of Computers, 33(3), 125-134.
[9] Zhou, W., Lian, J., Zhang, J., Mei, Z., Gao, Y., & Hui, G. (2023). Tomato storage quality predicting method based on portable electronic nose system combined with WOA-SVM model. Journal of Food Measurement and Characterization, 1-11.
[10] Cheng, X., Yang, H., Yuan, L., Lu, Y., Cao, C., & Wu, G. (2022). Fault Feature Enhanced Extraction and Fault Diagnosis Method of Vibrating Screen Bearings. Machines, 10(11), 1007.
[11] Lian, J., Ma, L., Wu, X., Zhu, T., Liu, Q., Sun, Y., ... & Lou, X. (2023). Visualized pattern recognition optimization for apple mechanical damage by laser relaxation spectroscopy. International Journal of Food Properties, 26(1), 1566-1578.