Comparing the Accuracy of Prediction Models based on Ensemble Machine Learning Schemes

Carlos Hernández
Departamento de Ingeniería Industrial y Sistemas
Universidad de La Frontera
Temuco, Chile
carlosalberto.hernandez@ufrontera.cl

Álvaro Alvar
Instituto de Ingeniería Industrial y Sistemas
Universidad Austral de Chile
Valdivia, Chile
alvaro.alvar@alumnos.uach.cl

Abstract

This research analyzes the influence of the configuration of ensemble learning algorithms' accuracy when predicting the annual production of honey for export in the south of Chile. The research is carried out following a classic 4-stage methodology (analysis, design, development, and validation). During the analysis, data is gathered and preprocessed. During the design, independent variables, ensemble algorithms, and performance metrics (correlation coefficient, MAE and RMSE) are defined. Construction and validation are carried out using the software WEKA. To build the models, 9 variables are considered. The dataset is split up in a subset for training and test (80%) and another one for validation (20%). The predictions are obtained by means of configuring a stacking scheme as ensemble and interchanging a support vector machine, a linear regression, a decision tree, and a Gaussian process as meta or base learners. According to the results, while the correlation coefficient between predictions and actual values fluctuates significantly in the range of 18% to 46%, MAE does it between 32% and 37%. In conclusion, although being inaccurate, results suggests that the arrangement of the meta and base algorithms within the ensemble does affect the prediction accuracy.

Keywords

Predictive Model, Machine Learning, Ensemble Algorithms, Stacking, Honey Production

1. Introduction

The decline in the population of honeybees is a matter of global concern. Despite such phenomenon, the consume of honey increases steadily year after year. In the global honey market, the most important producers and exporters are China and Argentine. And the mayor honey importers are United States and Germany. There are several factors that influence the production of honey (Rocha, 2017). Some of them are weather conditions, economic factors, and environment issues (Delgado et al., 2012). Several predictive models have been proposed in the past to predict the production of honey, including the aforementioned factors. From linear regressions to artificial neural networks (Çevrimli et al., 2020). This investigation, however, analyzes the influence of the configuration of an ensemble learning algorithm on the predictions´ accuracy using information about the annual production of honey for export in the southern regions of Chile.

For the purposes of this research historical data from the last 10 years are considered. The dataset is split up in a subset for training and test (80%) and another subset for validation (20%). The approach to predict the annual production of honey for export considers the development of several predictive models with different configurations. Each model consists of an ensemble scheme, a meta-algorithm and three base learners (Valentini and Masulli, 2002). The selected algorithms are linear regression, a Gaussian process regression (Wilson et al., 2011), support vector machine (Basak et al., 2007; Guenther and Schonlau, 2016; Rivas-Perea, 2013), and the decision tree M5. Since the yearly production

is a priori known, the problem can be classified as a supervised learning task. In all cases, instead of a simple hold-out, a 10-fold cross validation is applied. Only the dataset for training and test is used to build the predictive models. To evaluate their performance, three metrics are considered: correlation coefficient, MAE, and RMSE.
It is well documented in the literature that an ensemble algorithm usually improved the accuracy of the predictions made by the single algorithms that are part of it. However, how significant the improvement really is when working with small datasets and how the configuration of the ensemble affects the accuracy of the predictions are some of the questions that this investigation addresses.

1.1 Objective

To determine the influence of the configuration of an ensemble machine learning algorithm on the predictions' accuracy when working with small datasets by means of building and comparing predictive models to estimate the annual production of honey for export in the south of Chile.

2. Literature Review

2.1 Machine learning

Machine learning is usually defined as a branch of artificial intelligence (IA). It uses techniques and algorithms to find patterns and to learn from datasets through experience. The learning can be achieved in different ways. While supervised learning requires datasets that contain the desired class or value to be predicted, unsupervised learning the desired class is not known.

2.2 Hold out and cross-validation

In machine learning, holding out refers to the split up of a dataset into a subset for training and another for testing purposes. The idea using a test dataset is to assess the performance of the predictive model with unseen data. Usually, the split proportion is 80% for training and 20% for testing. On the other hand, cross-validation refers to the random split up of a dataset into k folds. During building, k-1 folds are used for training while the left one is used for testing. Training and testing are repeated iteratively k times until all folds have been used for training and for testing (Figure 1). When implementing a cross-validation, each iteration produces different results because the fold for training and for testing have been interchanged. The k results of a cross-validation are finally averaged (Powers and Atyabi, 2012).

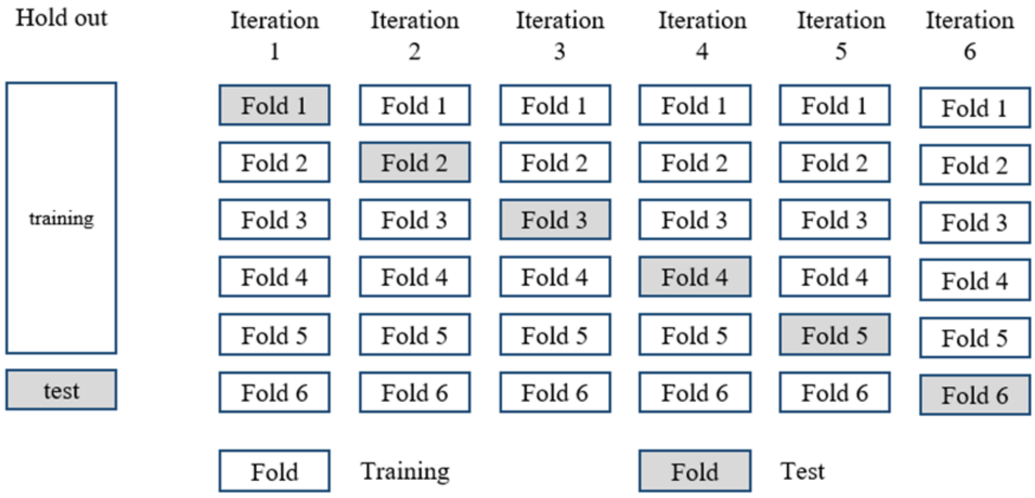

Figure 1. Cross-validation with k=6

2.3 Replication

A replication can be understood as the repetition of an experiment under similar conditions to estimate the variability of phenomenon under study. When using cross-validation, the dataset partition in k folds can be done in different ways, it all depends on a specific seed used for the partitioning. Thus, different seeds give rise to different folds and, most likely, to different results. By means of replicating the experiment, it possible to estimate the variation of those results.

2.4 Ensemble algorithms

Ensemble algorithms are characterized by a multi-echelon structure to carry out the learning task. In the simplest configuration there is a meta-algorithm (level 1) that learns from the predictions made by several base learners (level 0). An ensemble usually can predict better than any of its single algorithms. There are several ensemble schemes, being the stacking one of the most widely used (Figure 2).

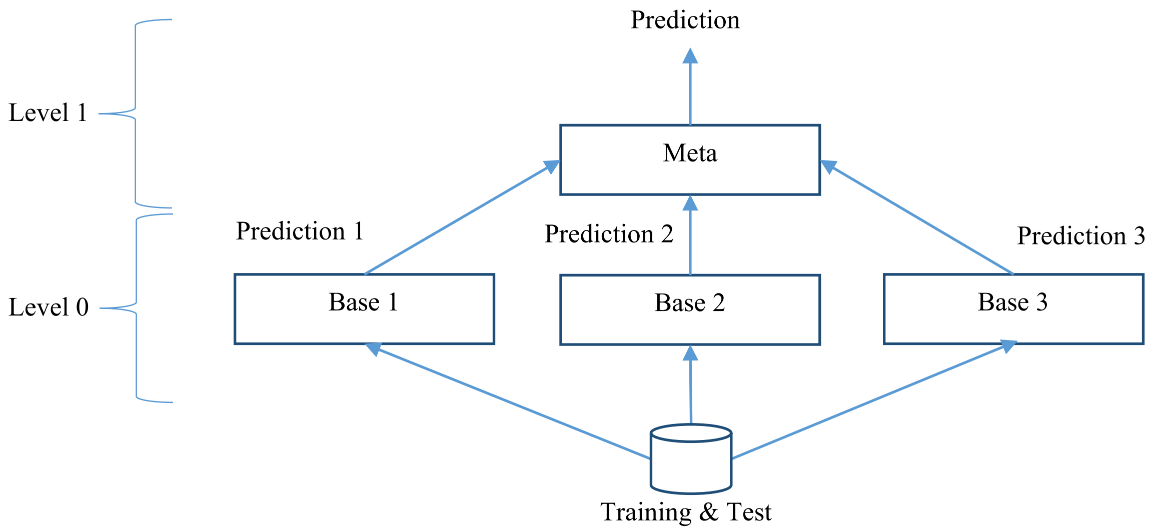

Figure 2. Ensemble algorithm configuration

3. Methods

The present research was carried out following a classic 4-stage model: analysis, design, construction, and validation (Figure 3).

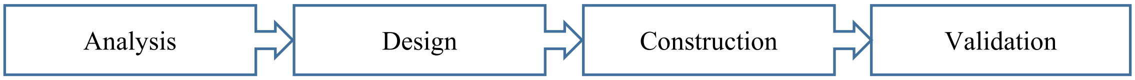

Figure 3. Four-stage model

3.1 Analysis

In the analysis, information about economy indicators, meteorological conditions, and environmental situation in Región de Los Ríos, southern Chile, between 2009 and 2018 is collected, preprocessed, and organized in a monthly basis. The scope of the investigation is defined at this point. Due to the lack of complete and accurate data, only 60 out of 120 months between 2009 and 2018 are finally considered in the research. Independent variables are grouped into several domains according to their nature. For practical reasons, some data are presented as a monthly average (e.g., temperature, relative humidity). With the exception of the annual production of honey, all the dependent variables are included in the models (Table 1).

Table 1. Domain and variables

	Variables (average monthly)	Economy	Meteorology	Environmental
A.1	Price per kilogram (USD)	√		
A.2	Temperature		√	
A.3	Relative humidity		√	
A.4	Dew point		√	

A.5	Wind speed		√	
A.6	Precipitation		√	
A.7	Particulate matter 2.5			√
A.8	Particulate matter 10			√
A.9	Honey annual production for export (ton)	√		

3.2 Design

The complete dataset, a matrix of 60 rows (instances) by 9 columns (attributes), is split up to create 2 subsets. The first one for training and test, contains 50 instances (80% of the data). The remaining 10 instances, 20% of the data, are left in a separate dataset to be used during the validation. An instance can be understood as a row containing all the actual values for every attribute in a single month.

The resulting models are compared by means of the corresponding correlation coefficient, MAE (mean absolute error), and RMSE (root mean squared error) (Weijie and Yanmin, 2018).

For the purposes of this investigation, based on preliminary trial and error experiments with the training and test data, four machine learning algorithms are selected. Namely, support vector machine, linear regression, decision tree, and Gaussian process regression. The ensemble scheme is the well-known stacking (Divina et al., 2018).

To identity the influence of ensembles' configuration on the prediction accuracy, different combinations are defined. Additionally, to visualize how individual learning algorithm performs, the first four models consist of single algorithms (Table 2).

Table 2. Predictive models' configurations

	Ensemble	Meta	Base 1	Base 2	Base 3
M1			Decision tree M5		
M2			Linear regression		
M3			Support vector machine		
M4			Gaussian process		
M5	Stacking	Decision tree	Linear regression	Support vector machine	Gaussian process
M6	Stacking	Linear regression	Decision tree	Support vector machine	Gaussian process
M7	Stacking	Support vector machine	Decision tree	Linear regression	Gaussian process
M8	Stacking	Gaussian process	Decision tree	Support vector machine	Linear regression

Although not explained in detail, the use of the linear regression scheme indicates that a correlation matrix analysis, R^2 adjusted numbers, and p-values are being considered. Even when all the attributes are initially considered to build a model, not all of them are finally expressed in the resulting mathematical formulation of the resulting model.

3.3 Construction

All models presented in this work are built using the software WEKA (Witten et al., 2017). The models are trained and tested applying a cross-validation scheme of k=10 folds on a dataset of 50 instances (Table 3).

Table 3. Models' performance with training and test data and cross-validation k=10 (50 instances)

Model	Correlation (%)	MAE (tons)	RMSE (tons)
M1	52.56	46.39	57.87
M2	51.92	46.71	58.48
M3	50.86	44.88	59.63
M4	50.93	45.12	57.61
M5	54.93	43.52	55.80
M6	55.33	43.16	55.69
M7	58.88	41.88	55.61
M8	31.08	55.63	65.82

It is well documented that cross-validation helps reduce the risk of overfitting. However, the influence of the folds partitioning remains. Replicating experiments is a simple way to minimize this problem by means of using a different fold partitioning in each replication. For the purposes of this research, 10 replications are run. This means that each model is trained and tested 100 times. Results show that the correlation coefficient fluctuates barely above 50% and its corresponding standard deviation are always between 31% and 35% (Table 4).

Table 4. Models' performance with training and test data, cross-validation (k=10) and 10 replications

	Correlation coefficient (%)		**MAE (tons)**		**RMSE (tons)**	
Model	**Mean**	**St. Deviation**	**Mean**	**St. Deviation**	**Mean**	**St. Deviation**
M1	49.50	35.50	48.38	15.40	57.44	16.34
M2	54.57	33.02	47.42	15.48	56.16	16.26
M3	54.15	33.50	46.33	16.62	57.13	19.50
M4	51.81	33.37	47.51	16.34	56.63	17.33
M5	51.16	33.99	47.87	16.32	57.11	17.84
M6	51.32	34.72	47.92	16.27	56.94	17.70
M7	54.12	31.65	46.61	15.40	56.28	17.36
M8	49.87	34.61	56.39	15.88	64.41	16.25

3.4 Validation

The models' validation is carried out using the validation dataset, which contains unseen 10 instances. This dataset is totally unknown to the predictive models.

4. Data Collection

The results of the validation stage reveal that only in two cases, M4 and M7, the correlation coefficients are consistent with those obtained with the training data. (Table 5).

Table 5. Models' performance with validation data (10 unseen instances)

Model	**Correlation (%)**	**MAE (tons)**	**RMSE (tons)**
M1	18.01	59.77	82.73
M2	41.24	49.44	71.29
M3	37.19	61.37	88.98
M4	46.08	49.13	69.85
M5	41.24	47.00	72.84
M6	41.24	47.08	73.02
M7	44.16	48.15	72.94
M8	40.39	48.74	78.05

5. Results and Discussion

The results show significant differences in the performance of the models, which indicates that the learning algorithms and the configuration of the ensemble are relevant. Among the ensemble-based models, only M7's predictions with unknown data seem to be consistent with the prediction from the training phase.

5.1 Numerical Results

Two of the most consistent models are M4 and M7. The comparison between their predictions shows correlation coefficients of 46% and 44%, and percentual MAEs of 37% and 32% respectively (Table 6).

Table 6. Actual values v/s predictions: M4 and M7 (10 unseen instances)

		Model 4 (MAE 37%)		**Model 7 (MAE 32%)**	
Instance	**Actual value (tons)**	**Prediction (tons)**	**Abs. Error (%)**	**Prediction (tons)**	**Abs. Error (%)**
1	105.00	148.30	41.24	136.43	29.93
2	105.00	90.04	14.25	63.71	39.32
3	105.00	115.20	9.71	99.13	5.59
4	42.32	86.29	103.90	69.58	64.41
5	160.80	145.34	9.61	156.24	2.84
6	80.40	115.47	43.62	102.58	27.59
7	201.00	109.30	45.63	91.09	54.68
8	321.60	139.21	56.71	132.73	58.73
9	100.50	123.51	22.90	115.57	15.00
10	140.70	109.36	22.27	105.67	24.90

6. Conclusion

Since the performance of predictive models based on machine learning algorithms is the result of a learning process, supervised in this case, having a complete and large enough dataset is important. Small datasets, like the one studied in this work, might lead to an incomplete learning process and consequently to inaccurate predictions. Even though, ensemble schemes are rather complex, the results are not much better than those produced by single algorithms. This is a good example that not always an ensemble can overcome the lack of an adequate dataset. Although cross-validation might help get rid of the partitioning effect that affects the simple hold out, the results in Table 4 show that when the dataset is small (50 instances) the selection of a large k (e.g., k=10) might produce folds so small that testing results varies significantly. In this case, while the mean of the correlation coefficient is close to 50%, its standard deviation is in the range of 30%.

Despite of being minimal, the implementation of an ensemble with four algorithms does produce more consistent results than learning schemes acting alone as shown in Table 5 for models 5, 6, 7, and 8. Finally, based on the results, it can be concluded that more prediction consistency is achieved when working with ensemble schemes. However, their capabilities cannot overcome the lack of an adequate dataset.

References

Basak, D., Pal, S., and Patranabis, D., Support Vector Regression, *Neural Information Processing – Letters and Reviews*, vol. 11, no. 10, 2007.

Çevrimli, M., Arikan, M., and Tekindal, M., Honey price estimation for the future in Turkey; example of 2019- 2020, *Ankara Üniv Vet Fak Derg*, vol. 67, pp. 143-152, 2020.

Delgado, D., Pérez, M., Galindo-Cardona, A., Giray, T., and Restrepo, C., Forecasting the Influence of Climate Change on Agroecosystem Services: Potential Impacts on Honey Yields in a Small-Island Developing State, Hindawi Publishing Corporation, 2012

Divina, F., Gilson, A., Gómez-Vela, F., Garcia, M., and Torres, J., Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting, *Energies 2018*, 11, 949, 2018.

Guenther, N., Schonlau, M., Support vector machines, *The Stata Journal*, vol. 16, no. 4, pp. 917–937, 2016.

Powers, D., Atyabi, A., The Problem of Cross-Validation: Averaging and Bias, Repetition and Significance. *2012 Spring World Congress on Engineering and Technology*, SCET 2012 - Proceedings. 1-5, 2012.

Rivas-Perea, P., Cota-Ruiz, J., Garcia, D., Perez, J., Venzor, Quezada, A., and Rosiles, J., Support Vector Machines for Regression: A Succinct Review of Large-Scale and Linear Programming Formulations, *International Journal of Intelligence Science,* vol. 3, pp. 5-14, 2013.

Rocha, H., Dias, J., Honey Yield Forecast Using Radial Basis Functions, *International Workshop on Machine Learning, Optimization, and Big Data Machine Learning, Optimization, and Big Data: Third International Workshop*, Volterra, Italy, pp. 483-495, 2017.

Valentini, G., Masulli, F., Ensembles of Learning Machines, *Conference: Neural Nets, 13th Italian Workshop on Neural Nets*, Vietri sul Mare, Italy, 2012.

Weijie, W., Yanmin, L., Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model, *IOP Conference Series: Materials Science and Engineering*, 2018.

Wilson, A., Knowlesy, D., and Ghahramaniz, Z., Gaussian Process Regression Networks, arXiv:1110.4411v1 [stat.ML], 2011.

Witten, I., Frank, E., Hall, M., and Pal, C., *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Edition, Morgan Kaufmann, Cambridge, 2017.

Biographies

Carlos Hernández is an industrial engineer, consultant, and university professor. He earned Licentiate Degree in Engineering from Universidad de La Frontera, Temuco, Chile, Master of Sciences in Computational Engineering and Doctor of Engineering from Technische Universität Braunschweig, Brunswick, Germany. He is the author of several scientific and engineering articles. He has taught lectures in Discrete Event Simulation, Supply Chain Management, Engineering Economics, Corporate Finances, Financial Engineering, Business Analytics, Data Mining and Machine Learning for engineering students. He has developed a professional career working for large multinational companies (PricewaterhouseCoopers, BHP Billiton, and Merck Sharp & Dohme). He also worked as a scientific researcher in the Institut für Produktionsmesstechnick at TU Braunschweig, Germany. His research interests include manufacturing process simulation, transportation systems simulation, supply chain design and simulation, and machine learning for finances. He is a member of IEOM.

Álvaro Alvar is an industrial engineer. He earned a B.S. and Licentiate Degree in Engineering from the Universidad Austral de Chile, Valdivia, Chile. His research interests include machine learning and forecasting methods.