Application of wrapper methods for feature selection in modelling ripening process of a viticulture crop

ROBERTO FERNANDEZ-MARTINEZ, JULIO FERNANDEZ-CENICEROS, ANDRES SANZ-GARCIA, RUBEN LOSTADO-LORZA, FRANCISCO JAVIER MARTINEZDEPISON-ASCACIBAR
Project Management Area. EDMANS research group (http://api.unirioja.es/edmans/)
Department of Mechanical Engineering
University of La Rioja
Luis de Ulloa 20. 26004. Logroño. La Rioja
SPAIN
roberto.fernandez@unirioja.es

Abstract: - The importance of final quality in agricultural products is growing ever greater, and is causing many crops to be monitored. This trend and the improvement of devices that can obtain the necessary information, make that it is stored large amount of information to work with. Apply learning algorithms in this amount of variables and data stored does that it is returned high execution times and calculation errors. To solve this problem, it is used feature reduction methods to select the most relevant information in order to reduce execution times and errors generated. Specifically, in this study, wrapper methods are used to select the most influential environmental variables during viticulture crop ripening.

Key-Words: - wrapper methods, feature selection, ripening grape berries, viticulture crop.

1 Introduction

Agricultural monitoring systems gather crowds of physical variables during the growing and ripening season of crops, especially now that new technologies give the possibility to have sensors that can obtain data from almost any physical variable in order to be measured [11] [22]. This means that it is possible to perform a collection of information about what is happening at all development stages. And thanks to a forthcoming study of collected data, it is possible to extract previously unknown knowledge to help to improve the process and get a better crop management based in decision support systems [3].

With all this amount of data, classified by variables, it is performed a study trying to obtain information to assist farmers in order to improve the process [28]. These studies are conducted with data mining techniques that turn data into information that helps to control crop [4] [29] [32]. All these techniques are one of the research areas where more progress is being made, although it should be noted that many of data collected may not have useful information for this study purpose. And when the amount of data is extremely high is very complicated to detect which is the data that do not contain useful information and which does.

To solve this problem it is worked on algorithms that perform a selection of the most interesting variables for the selected goal [1] [23] [24]. The gathered data can be studied with these algorithms and determine which variables have the most significant information and which provide little or nothing.

This kind of algorithms is used in the monitoring of a crop over several productive seasons, to get the variables that have some influence on the cultivation. A large number of variables are associated with weather conditions, as they have significant influence on most crops. Although not all variables that can be collected by weather stations, which are now on the market, are useful to show a correlation with the crop evolution. So they are selected variables that can actually have meaningful correlations between measures from weather stations and information collected by farmers about the crop.

Using learning algorithms is possible to work with several methods that can be used to relate all this information, but when the amount of information is high, worse conclusions are obtained in many times. Besides computational cost is increased according the amount of variables and data, that is involved in the process.

To improve the efficiency of these methods is carried out a previous step with data. It is made a feature selection that allows reducing the execution time and simplifying the calculation process when applying learning algorithms. It is developed a reduction of variables and data gathered from weather stations, in

ISBN: 978-960-474-281-3

order to obtain a better correlation of these variables with those obtained in the process of crop maturation.

2 Problem Formulation

It is worked in the prediction of some physical and chemical variables (weight, sugar, acidity, ...) to learn how those features evolve during the maturation process in a vineyard and how they are influenced by different environmental parameters. The data used in this study were collected from some different study areas of La Rioja (Spain). Once collected, it is generated all the necessary variables to control this process, according to the recommendations of several authors [5] [12] [13] [26]. Concluding that the needed variables used to predict these features are 29. Furthermore, it is had a great amount of data of each of these variables, because the study is conducted with data collected over 7 years.

The work done with all these data generates several problems that thanks to the feature selection algorithms can be resolved [21]:

- The results quality is worse, because the initial data can have many variables, and some of them can confuse the algorithms with irrelevant information, that makes the algorithms work wrongly.
- Irrelevant variables and redundant information, that may not provide an increase in the information quality available within of each class, make to increase the learning time used by the algorithms
- The predictive accuracy is much worse since these data make the calculations generating some higher errors

3 Problem Solution

To solve the problem deriving from working with so many variables there are two methods that allow reducing this quantity. Feature transformation and feature selection:

- Feature transformation: According to this method, new variables are created from a transformation of the initial group. These transformations are made from linear or nonlinear combinations in order to reduce the dimensionality of the dataset and lose as little information as possible [20].
- Feature selection. However, according to this other method, the variables are not transformed but the most significant of all are chosen to make up the database [6] [15] [9]. Feature selection consists of selecting what type of characteristics or traits are the best suited to describe the variables that it is wanted to predict. In order to do this, it is located the features that affect the problem in a more crucial way.

Of these two methods to reduce variables, it is considered in this case, that the second is more useful for researchers because it does not transform the variables and shows more clearly how each one affects to the final result. The advantages generated by this method are:

- Improved compression of the generated models, because the models are composed of fewer variables.
- Reduction of computational cost when the final model is generated. Especially in more complex models.
- Improved accuracy of final model, because reduction of non-significant variables improves its accuracy.

3.1 Feature selection methods: filters and wrappers

There are two main groups of feature selection methods. One called filter or indirect approach, and another one called wrapper or direct approach.

3.1.1 Filter

These methods make use of heuristic algorithms to determine the optimal subset of features. Where the final solution is not determined directly but it is got making attempts. To obtain it, a group of possible solutions are generated during the process according to a given pattern. These solutions are tested using the criteria that characterize the solution. And of all the solutions generated, invalid solutions are not taken into account [16] [18].

The main advantage of this algorithm is its faster speed in the calculation and the computational cost savings.

3.1.2 Wrapper

In this method, each data subset is evaluated by means of learning algorithms. This gives greater accuracy, but also carries a higher computational cost [17].

Decision trees are often used to examine the attributes that are not used or used in fewer rules.

3.2 Used methods

On many occasions, filters methods are most commonly used for its speed and acceptable results. There are even a greater number of techniques based on these methods. But in this study it is studied wrapper methods of subset selection type, as it is considered they are the ones that produce more accurate results. Because it is considered acceptable to have a higher duration and a higher computational cost in his generation.

Wrappers methods used to carry out this study have been implemented within the WEKA tool and are:

- ClassifierSubsetEval [31]: Evaluates subsets of training data attributes or a set of independent test, using classifiers.
- WrapperSubsetEval [17]: Evaluates subsets of attributes using classifiers. Internal cross validation is used to estimate the accuracy of the learning system in each set. And later assign a merit to each of given groups of variables. It is also developed several trees to evaluate each data set to get merit estimation with less error.

Several classifiers are used with different learning strategies, for each of the selected methods:

- Linear Regresion (LR) [30].
- Multilayer Perceptron (MLP) [10].
- Decision stump (DS) [2].
- M5P (M5P) [25].
- RepTree (REP) [31].

3.3 Search methods

Like feature selection methods, it is unclear which search method is the most appropriate for each case.

Table 1. Initial variables group used to put into practice feature selection methods.

Vineyard variables	
Variety	Var
Vineyard age (year)	Age
Altitude (m)	Altit
Environmental variables related to the amount of rainfall	
Total rainfall over the preceding week (mm.)	RFW
Total rainfall over the preceding two weeks (mm)	RF2W
Total rainfall over the preceding three weeks (mm)	RF3W
Total rainfall since the beginning of the year (mm)	RFY
Total rainfall since bud break (mm)	RFBB
Total rainfall during the penultimate week (mm)	RFW2
Total rainfall during the penultimate and antepenultimate week (mm)	RF2W2
Total rainfall between bud break and flowering (mm)	RFBBF
Total rainfall between flowering and setting (mm)	RFFS
Total rainfall between setting and véraison (mm)	RFSV
Total rainfall between véraison and harvest (mm)	RFVH
Environmental variables related to wind, humidity and weight	
Prevailing wind direction over the preceding week (N,S,E,W)	Dir
Average relative humidity over the preceding week (%)	Hum
Minimum relative humidity over the preceding week (%)	HumMin
Maximum relative humidity over the preceding week (%)	HumMax
Average wind speed in Km/h over the preceding week (Km/h)	Speed
Maximum wind speed in Km/h over the preceding week (Km/h)	SpeedMax
Weight of 100 berries	W100B
Environmental variables related to temperature	
Average temperature over the preceding week (ºC)	Temp
Minimum temperature over the preceding week (ºC)	TempMin
Maximum temperature over the preceding week (ºC)	TempMax
Aggregate of average daily temperatures since the beginning of the year (ºC)	STemp
Days with maximum temperatures above 40º C	D40
Days with average temperatures above 18º C during maturation	DM18
Days with maximum temperatures above 30º C during maturation	DM30
Average differences between maximum and minimum daily temperature during maturation (ºC)	DDN

And so that several studies, comparing different search algorithms are made by several authors [14] [19], although in this study is only used the following methods.

- Best first: Greedy hill-climbing with backtracking [27].
- Genetic search: Search using a simple genetic algorithm [7].
- Linear forward selection: Extension of best first [8].

3.4 Results

The variables selected initially are shown in Table 1. And from a prescribed number of variables, this study shows that as a preliminary step to calibrate a model, there are several techniques for reducing the number of variables.

In Tables 2 and 3 are shown, for different search methods and wrappers methods used, which are the variables selected. To see how feature selection affects to the problem, a study is made of how much influence has a model calibration with all variables or just using the variables selected by the feature selection algorithms. The finally selected variables were the ones which have shown best figures with all used methods.

In this case, the variables that result more suitable according to the algorithms are: Var, Age, Altit, Hum, HumMin, HumMax, SpeedMax, RFY, RFBB, RFBBF, RFSV, TempMin, STemp, DM18 and DM30. The accumulated percentages of all the variables, from which these features are selected, are shown in Table 4.

A test is done to develop a regression model of sugar concentration that owns a grain of grape of a vineyard during its maturation, in order to verify that the work carried out with the feature selection methods is useful. The regression model chosen is a neuronal network formed by five neurons. This model is calibrated and tested, as much with all the initial data, like just by the variables selected by the methods used in this experiment. Not only it contributes to diminish the time used in the model generation but it improves the forecast results of this regression (Table 5).

The studied error indices are the following:

- Correlation coefficient (CORR)
- Root mean squared error (RMSE)
- Mean absolute error (MAE)
- Root relative squared error (%) (RRSE)
- Relative absolute error (%) (RAE)

Also it is possible to observe like comparing the results obtained in the correlation between the real data and the calculated, it is obtained better results using only the selected variables (Fig. 1).

Table 2. Results of applying the method ClassifierSubSetEval.

Variables	ClassifierSubSetEval														
	Best First					Genetic search					Linear Forward selection				
	DS	LR	M5P	MLP	REP	DS	LR	M5P	MLP	REP	DS	LR	M5P	MLP	REP
Var	90	100	40	80		90	100	70	100		90	100	40	80	
Age		100	100	70	80	20	100	80	80	100		100	100	70	80
Altit	10	100	10	80	100	40	100	40	100	100	10	100	10	80	100
Dir		30	80	40	60	20	10	90	60	30		30	80	40	60
Hum		100	100	50	100	10	100	100	80	90		100	100	50	100
HumMin		100	100	80	100		100	90	70	90		100	100	80	100
HumMax		100	100	80	90		100	100	80	90		100	100	80	90
Speed			100	50	80		10	70	80	80			100	50	80
SpeedMax		100	100	70	70	20	100	100	70	50		100	100	70	70
W100B			20	80		20		30	100				20	80	
RFW		60	80	80	30	20	30	50	60	60		60	80	80	30
RF2W		30	90	60	30	40	50	60	20	20		30	90	60	30
RF3W		60	70	80	50	10	30	60	70	60		60	70	70	50
RFY		40	80	50		30	50	100	60	60		40	80	60	
RFBB		100	100	70	100	20	100	90	70	60		100	100	70	100
RFW2		20	60	40	30	30	70	60	40	30		20	60	40	30
RF2W2		10	20	30	60	10	20	50	30	60		10	20	30	60
RFBBF	60	100	60	80	20	90	80	80	80	60	60	100	60	80	20
RFFS			70	80	40	20	50	50	70	70			70	80	40
RFSV		100	90	80	50	10	100	90	80	50		100	90	80	50
RFVH	40	80	30	30	80	40	70	50	60	90	40	80	30	30	80
Temp		100	20	60	70	40	100	50	60	30		100	20	60	70
TempMin		100	80	60	40	50	100	100	60	40		100	80	60	40
TempMax		100	60	60	30	50	100	30	70	50		100	60	60	30
STemp	100	20	100	60	80	100		100	70	60	100	20	100	60	80
D40		100	60	30	50		100	60	20	30		100	60	30	50
DM18		100	90	90	40		100	100	60	60		100	90	90	40
DM30		100	100	60	30		100	100	50	40		100	100	60	30
DDN		100	20	60	20		100	40	70	40		100	20	60	20

Table 3. Results of applying the method WrapperSubSetEval.

Variables	WrapperSubSetEval														
	Best First					Genetic search					Linear Forward selection				
	DS	LR	M5P	MLP	REP	DS	LR	M5P	MLP	REP	DS	LR	M5P	MLP	REP
Var	30	100	20	100		30	100	40	100		30	100	20	100	
Age		50	40	100		30	50	70	90			50	40	100	
Altit	70	100		100		70	100	30	100		70	100		100	
Dir			80	70	70	10		60	60	60			80	70	70
Hum		100	100	70	50		100	100	70	50		100	100	70	50
HumMin		50	100	100	70		50	100	100	70		50	100	100	70
HumMax		100	100	90	30	10	100	100	90	60		100	100	90	30
Speed		50	90	70	80		50	60	80	70		50	90	70	80
SpeedMax		100	100	100	50	10	100	100	90	40		100	100	100	50
W100B		50		100			60	30	80			60		100	
RFW		60	60	60	40	10	50	50	60	20		60	60	60	40
RF2W		40	80	80	60	20	40	70	60	40		40	80	80	60
RF3W		10	30	80	60	20	10	70	80	50		10	30	80	60
RFY	20	90	90	90	80	30	60	100	100	70	20	90	90	90	80
RFBB		100	100	90	100	20	100	100	100	80		100	100	90	100
RFW2			70	30	30		10	40	30	40			70	30	30
RF2W2		10	40	40	20	10	40	10	90	30		10	40	40	20
RFBBF	30	100	100	90	90	30	90	70	100	90	30	100	100	90	90
RFFS		20	80	90	60	20		100	100	90		20	80	90	60
RFSV		100	100	90	100	10	100	90	100	80		100	100	90	100
RFVH	50	10	30	60	80	40	50	10	70	60	50	10	30	60	80
Temp		100	60	40	20	20	100	60	70	60		100	60	40	20
TempMin		100	70	90	50	20	100	90	90	90		100	70	90	50
TempMax		100	90	50	40	40	100	70	50	70		100	90	50	40
STemp	100	10	100	100	100	100	10	100	100	100	100	10	100	100	100
D40		80	60	40	40		80	50	60	50		80	60	40	40
DM18		100	90	100	70		100	70	80	70		100	90	100	70
DM30		100	100	70	60		100	100	90	100		100	100	70	60
DDN		100	30	70	10		100	40	40	40		100	30	70	10

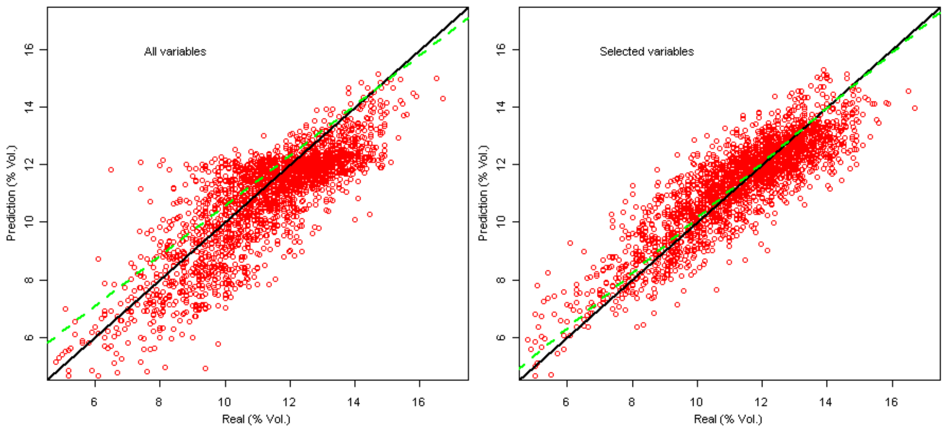

Figure 1. Existing correlation between the real data and the predicted ones. For all the variables and only the selected variables.

Table 4. Accumulated percentages of used methods.

Vineyard variables	
Var	1750
Age	1700
Altit	1820
Environmental variables related to the amount of rainfall	
RFW	1350
RF2W	1360
RF3W	1330
RFY	1750
RFBB	2260
RFW2	910
RF2W2	810
RFBBF	2230
RFFS	1450
RFSV	2130
RFVH	1520
Environmental variables related to wind, humidity and weight	
Dir	1260
Hum	2040
HumMin	2070
HumMax	2110
Speed	1540
SpeedMax	2060
W100B	830
Environmental variables related to temperature	
Temp	1530
TempMin	1920
TempMax	1690
STemp	2280
D40	1370
DM18	2000
DM30	1920
DDN	1290

Table 5. Errors observed when validating the model using all the variables and only the selected ones.

	CORR	RMSE	MAE	RRSE (%)	RAE (%)
All variables	0.7986	0.1016	0.0791	62.8631	61.2768
Selected variables	0.8723	0.079	0.0621	48.9106	48.139

4 Conclusions

It is observed that the use of feature selection algorithms is effective since when reducing the input variables, the understanding of the generated models improves. And since the model is compound of less number of variables, the problem is defined by the most significant variables in a clearer way. And as it is not used feature transformation methods these variables are more understandable.

Also a reduction of computational cost is verified when generating the final model. Mainly in the most complex models, since he is not the same to calibrate models with many variables that do not contribute significant information, that to calibrate only with the most significant. In the studied case the time of operation is reduced in a 34%.

In addition the precision of the final model has improved. The reduction of no significant variables improves the precision in the five studied indices to verify the error.

It is also verified, that all the used algorithms to reduce features do not get the same conclusions, although realising an analysis of all the methods jointly allow us to obtain the best solution for the problem resolution.

As final conclusion, it is determined that the application of these algorithms is useful with this kind of data and is advisable for future works.

Acknowledgment

The authors thank the University of La Rioja through FPI fellowships and the Autonomous Government of La Rioja for its support through the FOMENTA program 2010/13.

References:

[1] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning,

Artificial Intelligence, Vol. 97, 1997, pp. 245-271.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Chapman & Hall, New York, 1984.

[3] R.A. Cardenas Tamayo, M.G. Lugo Ibarra, J.A. Garcia Macias, Better crop management with decision support systems based on wireless sensor networks, *Electrical Engineering Computing Science and Automatic Control (CCE),* 2010, pp. 412-417.

[4] A. Ceglar, Z. Crepinsek, L. Kajfez-Bogataj, T. Pogacar, The simulation of phenological development in dynamic crop model: The Bayesian comparison of different methods, *Agricultural and Forest Meteorology*, Vol. 151, 2011, pp. 101-115.

[5] B.G. Coombe, Research on Development and Ripening of the Grape Berry, *Am. J. Enol. Vitic.*, Vol. 43, 1992, pp. 101-110.

[6] M. Dash, H. Liu, Feature Selection for Classification, *Intelligent Data Analysis,* Vol. 1, 1997, pp. 131-156.

[7] D.E. Goldberg, *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley Pub. Co., Reading, Mass, 1989.

[8] M. Gutlein, E. Frank, M. Hall, A. Karwath, Large-scale attribute selection using wrappers, *Computational Intelligence and Data Mining, CIDM '09,* 2009, pp. 332-339.

[9] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,* Vol. 3, 2003, pp. 1157-1182.

[10] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd edn, Prentice Hall PTR: Upper Saddle River, NJ, USA, 1999.

[11] J. Hwang, C. Shin, H. Yoe, Study on an Agricultural Environment Monitoring Server System using Wireless Sensor Networks, *Sensors,* Vol. 10, 2010, pp. 11189-11211.

[12] D.I. Jackson, P.B. Lombard, Environmental and Management Practices Affecting Grape Composition and Wine Quality - A Review, *Am. J. Enol. Vitic,* Vol. 44, 1993, pp. 409-430.

[13] R.S. Jackson, *Wine Science Principles and Applications,* Third Edition, Elsevier Inc, 2008.

[14] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, *Pattern Analysis and Machine Intelligence,* Vol. 19, 1997, pp. 153-158.

[15] G.H. John, R. Kohavi, K. Pfleger, Irrelevant feature and the subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning,* 1994, pp. 121-129.

[16] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, *Proceedings of the tenth national conference on Artificial intelligence*, 1992, pp. 129-134.

[17] R. Kohavi, G.H. John, Wrappers for feature subset selection, *Artif. Intell.* Vol. 97, 1997, pp. 273-324.

[18] D. Koller, M. Sahami, Toward Optimal Feature Selection, *Proceedings of ICML-96, 13th International Conference on Machine Learning,* 1996, pp. 284-292.

[19] M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition,* Vol. 33, 2000, pp. 25-41.

[20] H. Liu, H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining,* Kluwer Academic Publishers, 1998.

[21] H. Liu, H. Motoda, L. Yu, Feature Extraction, Selection, and Construction, *The Handbook of Data Mining.* Lawrence Erlbaum Associates, Inc. Publishers, 2003, pp. 409 - 423.

[22] F. Mazzetto, A. Calcante, A. Mena, A. Vercesi, Integration of optical and analogue sensors for monitoring canopy health and vigour in precision viticulture, *Precision Agriculture,* Vol. 11, 2010, pp. 636-649.

[23] Y. Peng, Z. Wu, J. Jiang, A novel feature selection approach for biomedical data classification, *Journal of Biomedical Informatics,* Vol. 43, 2010, pp. 15-23.

[24] S. Piramuthu, Evaluating feature selection methods for learning in data mining applications, *European Journal of Operational Research,* Vol. 156, 2004, pp. 483-494.

[25] J.R. Quinlan, Learning with continuous classes, *Proceedings of the 5th Australian joint Conference on Artificial Intelligence.* Singapore, 1995, pp. 343-348.

[26] P. Ribéreau-Gayon, D. Dubourdieu, B. Donèche, A. Lonvaud, Chapter 10. The Grape and its Maturation, *Handbook of Enology, Volume 1, 2nd Edition, The Microbiology of Wine and Vinifications,* John Wiley & Sons L, 2005.

[27] L. Rios, L. Chaimowicz, A Survey and Classification of A* Based Best-First Heuristic Search Algorithms, *Advances in Artificial Intelligence, SBIA 2010*, Springer Berlin / Heidelberg, 2011, pp. 253-262.

[28] T.R. Sinclair, N.G. Seligman, Crop Modeling: From Infancy to Maturity, *Agron. J.,* Vol 88, 1996, pp. 698-704.

[29] C.O. Stöckle, M. Donatelli, R. Nelson, CropSyst, a cropping systems simulation model, *European Journal of Agronomy,* Vol. 18, 2003, pp. 289-307.

[30] G.N. Wilkinson, C.E. Rogers, Symbolic description of factorial models for analysis of variance. *Journal of Applied Statistics,* Vol. 22, 1973, pp. 392-399.

[31] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition edn, Morgan Kaufmann, 2005.

[32] X. Zhang, M.A. Friedl, C.B. Schaaf, et al, Monitoring vegetation phenology using MODIS, *Remote Sensing of Environment,* Vol. 84, 2003, pp. 471-475.