Mutual Information-based Feature Selection in Spectrometric Data for Agriculture Applications

L.J. Herrera, V. Lafuente, R Ghinea, M.M Perez, I. Negueruela, H. Pomares, I. Rojas, A. Guillén

Abstract—**This work presents a feature selection comparative of Mutual Information-based feature selection methods in agriculture industry applications. The starting point is a set of datasets in which near infrared spectroscopy (NIR) were applied to a variety of peach, apple and two varieties of cherry used to predict fruit properties (firmness and soluble solid contents). LS-SVM regression was the method used for model assessment. The different MI-based feature selection methods used (5 in total), were compared by considering ROC curves for different number of features. This type of application is essential in constructing optimal applications and instrumentation in fruit harvesting time optimization in agriculture industry.**

Keywords: Feature selection, LS-SVM, Mutual Information, Agriculture Application, Vis/NIR

I. Introduction

Feature selection is a key preprocess step in any modeling problem. It consists on the identification of the truly relevant factors involved in the physical/chemical/natural system that is being modeled. The objective is also to select which of those are the most convenient ones to perform an accurate modeling of the problem [16].

Several methods exist for feature selection in the literature. One of the most well-known selection criteria is Mutual Information, coming from Shannon's Information Theory [14]. Others include criteria such as PCA, ICA, Delta Test, Gamma Test, and other correlation measures. This process is especially important in several industrial applications, in which selecting an optimal set of factors to model the processes is essential for good performance and cost, which are critical for the success and profitability of the business.

In food industry and agriculture one of the critical issues is the development of systems that allow determining certain internal parameters of the food without destroying it. In agriculture, a correct identification of the harvesting time is critical, and for that purpose, the internal parameters of the fruits have to be determined [5][7].

This work deals with the problem of determining the firmness and soluble solid contents (SSC) [8] of four types of fruit, the Prunus persica 'Calrico' peach, apple, and two varieties of cherry, Cashmere and Chelan, from Vis/NIR data. Later in a second stage, Vis/NIR data and acoustic data are together used to improve firmness prediction in Calrico peach. And for this purpose, five feature selection algorithms are tested and their performance is evaluated. The conclusion of this work leads to the fact that one of the feature selection algorithm, backwards Markov Blanket-based feature selection is clearly superior to the other forward and backward, well-known, feature selection algorithms tested. Secondly it is obtained that again with this algorithm, with only two Vis/NIR wavelengths plus the acoustic measure, performance in firmness in Calrico peach problem is optimal.

The rest of the paper is organized as follows. Section II describes the agriculture problem of fruit properties prediction in fruit. Section III introduces mutual information and the five variants of feature selection methods that will be compared in this work. Section IV briefly introduces Least Squares Support Vector Machines as the modelling methodology used for comparisons in this work. Section V presents and discusses the results obtained. Section VI exposes the conclusions drawn from this work.

II. Firmness and SSC prediction

Nowadays, consumers select the fruit not only for its appearance, but also for its internal quality. Due to this, it is necessary to develop systems that allow determining internal parameters of the fruit without destroying it. Soluble solid content and firmness are the two most widely used parameters to estimate fruit ripeness and quality.

Since any fruit is harvested it is separated from its source of nutrients. Nevertheless still from that moment their tissues breath and are physiologically active. Fruit ripening implies complex physical and chemical changes, such as softening, increased concentration soluble sugars, flavor and color changes. These processes are important because they influence changes that will occur during storage, transportation and commercialization and to some extent will affect its nutritional value and organoleptic characteristics [1]. Harvest date is not only useful in obtaining a quality product but also increases production and minimizes costs in agriculture [2].

This work has been supported by the GENIL-PYR-2014-12 project from the GENIL Program of the CEI BioTic, Granada, and the Junta de Andalucia Excellence Project P12-TIC-2082.

L. J. Herrera, H. Pomares, I. Rojas, A. Guillén are with the Computer Architecture and Computer Science Department of the University of Granada, 18071, Granada, Spain (e-mail: jherrera@ ugr.es).

V. Lafuente is with the Estación Experimental de Aula Dei (CSIC). Vegetal Nutrition Department, 50059, Zaragoza

R. Ghinea and M.M. Perez are with the Department of Optics, University of Granada, 18071, Granada, Spain

I. Negueruela is with the Department of Applied Physics, University of Zaragoza, 50013 Zaragoza.

Firmness is a parameter that is highly correlated with fruit ripeness and is ultimately very useful in determining the optimal harvest date. Firmness can be defined as the resistance to penetration force by the fruit pulp out [3]. Traditionally firmness was determined by the destructive 'Magness-Taylor' method (MT). The MT test involves attaching a dynamometer to a cylindrical rod, which is inserted 8 mm into the fruit pulp after having removed part of the skin. Although this test is inexpensive and fast, it destroys the fruit which limits its usefulness.

Sugars are the major soluble solids in fruit. Other soluble materials include organic and amino acids, soluble pectins, etc. Soluble solids are also measured by destroying the fruit as some juice has to be extracted and measured through refractometer or other instruments.

As firmness and SST measurements imply destructive techniques, they can only be applied to a rate of the pieces. Due to this, these techniques can not be applied for inline classification in the fruit centrals. Since 2003, non-destructive more practical and reproducible methods have been developed for estimating fruit firmness and SSC with the aim of replacing destructive ones, which moreover can be applied to all the fruit pieces, and not only to a selection of them.

The most well-known and emerging non-destructive technique currently used for fruit properties prediction is Near-Infrared Spectroscopy (NIRS). Many researchers are developing methods based on this technique to predict firmness and SSC values for different fruits. In the NIRS technique, a light beam strikes the fruit penetrating it a few millimetres. Part of this radiation is in the visible and infrared region.

Another non-destructive method involves the use of acoustic signals caused by vibrations or mechanical impacts to the fruit. The resonant frequency that is emitted by an object depends directly on its geometry, mass and modulus of elasticity from which the material is comprised [4]. Acoustical tests performed on the sample are stimulated by a low intensity impact, producing a vibratory response within audible range (20-20000 Hertz). The response is recorded with a microphone and the signal in time is processed using Fast Fourier Transform to obtain the corresponding signal in the frequency, which produces an acoustic firmness index [5][6].

Signals obtained in these or any other non-destructive method are related to the desired estimation parameter from which mathematical/computational models are obtained. Firmness MT obtained value is considered reference. Despite its variability, it is an acceptable reference in fieldwork. Nonlinear methods such as 'Least Squares Support Vector Machines' (LS-SVM)[10] have traditionally been used as regression method for this purpose, apart from traditional Partial Least Squares (PLS) method [9].

III. Mutual information feature selection methods

Mutual information is a non-linear correlation measurement from the Information Theory [14]. For two sets of continuous features, X and Y, it can be calculated by:

$$I(X,Y) = \int \frac{\mu_{X.Y}(x,y)\log(\mu_{X.Y}(x,y))}{\mu_X(x)\mu_Y(y)}dxdy$$

where $\mu_{X.Y}(x,y)$ is the joint probability density function (PDF) of X and Y, and $\mu_X(x)$ is the marginal density function of the set of features X. The advantage of this criterion over other correlation criteria, is that it is able to identify non-linear relations among the features involved.

Several attempts have been reported in recent literature for designing algorithms to identify the most relevant factors (wavelengths and/or other factors) for the prediction of chemical properties, many of those are based in mutual information [11][12][15]. Feature selection aims at identifying irrelevant and redundant features for their rejection. Identification of redundant features is critical in spectrometric problems, as nearby wavelengths provide usually similar information. The reduction in features needed to predict any magnitude is essential to reduce experimental and evaluation costs, but also for increasing the so-called generalization capability of the models, i.e., prediction capability on unseen data [16].

The MI estimator used in this paper is the nearest neighbors estimator extended from entropy estimation to the MI in [19]. Moreover, resampling methods according to [22] were used in order to strengthen the robustness of the mutual information estimation among the features.

The feature selection methods presented in this paper return a ranking of features in consecutive order of relevance in selecting a subset of features. The most relevant feature is in principle the optimal one for a single-feature subset; the two most relevant features are in principle the optimal ones for a two-features subset; the three most relevant features are in principle the optimal ones for a three-features subset of features; and so on. Thus, given the returned ranking, it is ideally expected that a subset of any size of the features selected according to this relevance ranking will provide better results than any other subset of features of the same size.

However, in order to provide a better insight of the real performance of the selected features in any problem, normally, a regression technique is used to evaluate different sizes of feature set (filter-wrapper approach) before selecting the optimal one.

The identification of the precise number of input features considered to perform the classification is normally performed by cross-validation (CV) evaluation of different regression models (one per each possible number of input features considered).

Performance in this work will be evaluated, due to the application involved, by identifying the best performance for different subset sizes in the first problem (Vis/NIR for SSC and Firmness prediction), and for the second problem (Vis/NIR + Acoustic measurement for Firmness prediction in Calrico peach), by identifying optimal number or features selected by each alternative of feature selection algorithm.

The five algorithms compared in this work will be two very successful and well-known Mutual Information-based algorithms appeared recently in the literature, both forward methods, the minimum Redundancy Maximum Relevancy algorithm (mRMR)[12], the Normalized Mutual Information Feature Selection method (NMIFS) [15], their respective backwards variants, and finally the Markov Blanket method taken from [18].

A. Markov Blanket MI Feature Selection

The proposed method is an approach first published by [17] and adapted for continuous features [18], which is based on the Markov blanket concept.

Given a set of input features X and an output feature Y, a set of features M_i in X is said to be a Markov blanket for a feature x_i in X with respect to Y, if $(I(\{M_i U\ x_i\}, Y) == I(x_i, Y))$, that is, if M_i has itself all the information that x_i has about Y. A Markov blanket is thus a group of features that subsumes the mutual information content in a certain feature, in practice and for our purposes, with respect to the objective feature.

The algorithm consists of a backwards feature selection method which starts with the complete set of features, and iteratively discards those which are detected to have a Markov Blanket in the remaining set X_G of features, i.e. those whose information with respect to Y is already present in the remaining set X_G of features.

The algorithm states the following steps:

1. Calculate the MI between every pair of input features $I(x_i, x_j)$
2. Starting from the complete set of input features $X_G = X$, iterate:
   a) For each feature xi, let the candidate Markov blanket M_i be the set of p features in X_G for which $I(x_i, x_j)$ is highest.
   b) Compute for each x_i
   $$Loss_i = I(\{M_i \cup x_i\}, Y) - I(M_i, Y)$$
   c) Choose the x_i for which $Loss_i$ is lowest and eliminate x_i from X_G.
3. Continue with step 2 until no features remain.

This way, a ranking of relevance of features (in reverse order) is obtained. Note that this way, features that have few influence with respect to the output feature (irrelevant features) will be soon discarded, as $Loss_i$ value should tend to 0. Similarly, redundant features will be iteratively discarded at earlier stages. Relevant features with low redundancy will be the last ones in being “chosen”. Further discussion about efficiency, character and operation of the algorithm can be found in [18]. The p value of the algorithm will take the value p = 1, as recommended in that work.

B. Minimum Redundancy Maximum Relevancy Feature Selection Algorithm

This algorithm was designed under the principle that directly selecting features according to mutual information has the problem of not considering the redundancy in MI that the input features can have among themselves. Thus, the mRMR (minimum redundancy - maximum relevancy) algorithm, proposed in [13], aims at a better identification of the relationships among the features. In this algorithm, the aim is to obtain a ranking of feature relevancy in incremental manner according to the following formulation, which starting from an empty set of features $S = \{\emptyset\}$, adds a new feature f_i to the current set which maximizes

$$G = I(c, f_i) - \frac{1}{\#S}\sum_{f_s \in S} I(f_i, f_s)$$

where #S is the current number of features in the selected set, C is the class feature and f_i, f_s are features of the problem.

C. Normalized Mutual Information Feature Selection Algorithm

Normalized Mutual Information Feature Selection (NMIFS) method [15] is a variant of the mRMR algorithm previously described, which uses the normalized MI measure by the maximum of the entropy of both sets of features:

$$NI(X, Y) = \frac{I(X, Y)}{\min\{H(X), H(Y)\}}$$

being H(X) and H(Y) the entropy of variables X and Y respectively.

The NMIFS method is also an iterative methodology that returns a relevance ranking of the input features with respect to the classification feature, aiming also to take to account not only their importance, but also the redundancy among themselves. Thus, starting from an empty set of features S = {∅}, the NMIFS algorithm iteratively selects the input feature f_i which maximizes:

$$G = I(C, f_i) - \frac{1}{\#S}\sum_{f_s \in S} NI(f_i, f_s)$$

where #S is the cardinality of the current selected set S.

IV. Least Squares Support Vector Machines

LS-SVMs are reformulations to standard SVMs, closely related to regularization networks and Gaussian processes but additionally emphasize and exploit primal-dual interpretations from optimization theory. LS-SVMs are a paradigm especially well suited for function approximation problems [13]. We avoid here further details on this methodology as we consider it is well-known already in the machine learning literature for regression problems.

Considering Gaussian kernels, the hyper-parameters of the model are σ_i as the width of the kernel, together with the regularization parameter γ. Hyper-parameters in LS-SVM were optimized using cross-validation and grid-search.

In order to reduce the computational cost of feature ranking evaluation, which requires training a different LSSVM model for each possible subset size, the Extreme Learning approach has been used [23]. It points out that it is possible to obtain successful classification results by using reasonable values of hyper-parameters. Then training a LS-SVM for the whole feature set, a pair of hyper-parameter values is obtained, which will be used for all possible feature subset size. This way, the computational cost of the evaluations is highly reduced.

ISBN: 978-988-19253-2-9 (revised on 23 September 2015) IMECS 2015
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. RESULTS

A. Data Description

As mentioned, four different fruit dataset were used in this work. First, the treated peach variety is of particular interest in the area of Bajo Aragón, which has applied for the certificate of origin 'Calanda'. Samples (260 fruits) were harvested in 2010 (150) and in 2011 (110). Apple '*Smoothee Golden Delicious'* dataset included 413 fruits, harvested in 2012, from May to October, in three steps, each 15 days. Prunus avium Cashmere and Chelan datasets included respectively 661 and 591 samples from two different seasons (2011 and 2012, during a period of two weeks). Non-destructive determinations (AWETA for Calrico and NIRS for all) were carried out prior to, destructive MT firmness and SSC determination.

Spectra from intact fruits were measured with a reflectance modular equipment Multispec instrument (AG Tec5, AM Frankfurt, Germany) equipped with a spectrometer SC - NEM I (Zeiss, Jena, Germany) (range: 400-1060 nm, $\Delta\lambda$ =1nm, 661 wavelengths in total). Acoustic firmness measurements were carried out by means of a commercial desktop acoustic firmness sensor (model AFS, AWETA, The Netherlands). The sensor recorded the weight and resonant frequencies of the acoustic vibration generated by gently tapping the fruit on the equatorial area, from which an acoustic firmness index was calculated.

Magness-Taylor firmness test were performed using a hand-held penetrometer Fruit Pressure Tester FT32 (Istituto per la Valorizzazione dei Prodotti Agricoli, Italy) with an 8 mm diameter probe. Fruit skin was removed with a blade at two positions around the equator and firmness measured. The firmness was measured in the same area where NIR reflectance spectra were acquired.

All data used in this work was obtained and prepared for thesis dissertation of V. Lafuente.

B. Simulation conditions

A Condor-based queue system in a 8-core PC, using MATLAB as programming language was used for the simulations. All simulations were formed by a twofold process for each problem, and each training-test subdivision in the cross-validation process for each problem. First, included obtaining optimal hyper-parameters for the LS-SVM machines, and also, obtaining the variable rankings for the five feature selection algorithms. Later, in a second stage and from the information obtained, a feature evaluation process was performed for each combination of features demanded in the comparisons, using the Extreme Learning approach.

C. Firmness and SSC prediction in fruit from Vis/NIR

In total 9 different problems were proposed from the given data: prediction of firmness from Vis/NIR data for the four fruits; prediction of SSC from Vis/NIR data for the four fruits; prediction of acoustic measurement (AWETA) from Vis/NIR for Calrico. 661 input variables and one output variables in all cases. From them, six problems were finally used to compare the five algorithms on these datasets, which were those whose estimated R2 values for the whole dataset was above 0.5 (see Table1).

TABLE I
10-FOLD CROSS VALIDATION PEROFRMANCE FOR THE WHOLE NIR FEATURE SET

Apple Firmness
Apple SSC
Calrico Firmness
Calrico Acoustic measurement
Cashmere SSC
Chelan SSC

The assessment over those 6 problems was done through 5-fold cross-validation, this procedure was performed to estimate the performance of the compared algorithms. I.e. the complete dataset was divided in five parts, using four of them to obtain the ranking of variables and estimate their performance through LS-SVM in the remaining part. R2 test training and test values reflect the mean over the 5 different cross-validation runs.

The number of variables used for algorithm comparisons were 1 (first variable selected by the algorithm), 2 (two first variables selected by the algorithm), 5, 10, 50, 100 and complete dataset (which could include slightly different performances using LS-SVM due to the way the data was organized in the n-dimensional space depending on the ranking obtained).

Table 2 show the mean test R2 cross-validation values for the six problems, the five algorithms, and the subset sizes taken. Bold values point out highest value in each case.

Results show superiority in the Markov Blanket and the Forward mRMR algorithms, with 22 highest values among all for the first algorithm and 19 for the second. However showing a larger superiority of the Markov Blanket algorithm for small subset sizes (but for one variable, i.e., subset sizes equal to 2, 5, 10 and 50). This result points out the convenience of Markov Blanket algorithm in the application of feature selection in this type of problems, it which it is desired a good performance with few wavelengths [25].

D. Firmness prediction in Calrico Peach from Vis/NIR and Acoustic measurement

For this more applied specific problem, 661 Vis/NIR input wavelengths, and one acoustic measurement (AWETA) were available to predict firmness in Calrico Peach. A 10-fold cross-validation procedure was used for evaluation; this CV division was also used for feature selection assessment. Results shown here partially extend those presented in [24].

Figure 1 shows the performance for one of the 10-fold CV (mean training and test R2 values) executions for the five feature selection algorithms tested for the peach problem for a selected range of features.

It can be observed from the training CV error (Fig 1.a)), that a suboptimum subset of variables is obtained for part of the algorithms for three variables, while other reach their first optimum for 5 or six variables. This is directly reflected in the test error performance. It is observed however for this execution, that the 3-size subset of variables selected by the Markov Blanket algorithm obtains optimal performance in comparison with the other algorithms.

Figure 2 moreover shows the mean performance of the five algorithms for the 10-fold cross validation execution. It is confirmed in this figure the superiority in the feature selection process by the Markov Blanket algorithm.

From the three variables selected, one is the AWETA measure and the other two selected wavelengths in all the groups were close to each other, surrounding the chlorophyll absorption region. This supports the use of this area of the Vis/NIR spectra to determine MT firmness. Similar results were obtained for other fruits [20][21].

In relation to the performance comparison between selecting 3 features and the whole subset, mean performances obtained R2 equal to 0.71 for three features, and 0.74 with the whole set of features. This difference however is low in terms of number of features.

It is to be claimed that there is the possibility to calculate the estimated state of ripeness with much simpler instrumentation which only requires two wavelengths in the visible region and acoustic measurement equipment.

VI. CONCLUSIONS AND FURTHER WORK

This work presented a feature selection comparative of Mutual Information-based feature selection methods in an agriculture industry application. Five different MI-based feature selection methods were compared by considering two different problems: first, four different fruit Vis/NIR datasets to predict firmness and SSC; second, Vis/NIR + acoustic measurement to predict firmness in Calrico Peach. First problem was evaluated using different feature subset sizes and their performance. Second problem was evaluated using the optimal number of features selected and their performance for firmness prediction. The best algorithm showed to be the Markov Blanket feature selection algorithm, obtaining a general better performance when selecting a low number of wavelengths in the first problem, and for the, more specific, second problem obtaining optimal performance with only 3 features from the 662 available.

Results shown in this work motivate the study over a larger database and probably using more feature selection alternatives in the purpose of identifying the algorithms that provide optimal performance in this general type of problems, i.e., properties estimation from spectrometric data on agriculture and food industry. These results are essential in the experimentation and design of optimal instrumentation for ripeness estimation in this type of industry.

REFERENCES

[1] Kader, AA, Flavor quality of fruits and vegetables. J .Sci. Food Agric. 88: 1863-1868 (2008)

[2] Thompson JF, Sistemas de cosecha. In Tecnologia Postcosecha de Productos Hortofrutícolas. 3ª edición. Kader, A. (Eds.). University of California, Division of Agriculture and Natural Resources, Oakland, California, USA 73-76 (2007)

[3] Mohsenin N N, Physical properties of plant and animal materials: structure, physical, characteristics and mechanical properties. New York: Gordon and Breach Science Publishers 664 (1986)

[4] Studman C J, Computers and electronics in postharvest technology: a review. Comp. Elec. Agricul 30: 109-124 (2001)

[5] Alves de Oliveira, G., Bureau, S., Renard, C.M., Pereira-Netto, A., Casthilos, F.Comparision of NIRS approach for prediction of internal quality traits in three fruit species. Food Chemistry, 143, 223-230 (2014)

[6] Valente M, Leardi R, Self G, Luciano G and Pain JP, Multivariate calibration of mango firmness using vis/NIR spectroscopy and acoustic impulse method. J Food Eng 94:7–13 (2009).

[7] Pissard, A., Fernández-Pierna, J., Baeten, V., Sinnaeve, G., Lognay, G., Mouteau, A., Dupont, P., Rondia, A., Lateur, M. Non-destructive measurement of vitamin C, total polyphenol and sugar content in apples using near-infrared spectroscopy. J Sci Food Agric.jsfa 5779 (2012)

[8] Li J, Huang W, Zhao C, Zhang B, A comparative study for the quantitative determination of soluble solids content, pH and firmness of pears by Vis/NIR spectroscopy. J. Food Eng.16:324-332 (2013)

[9] Travers, S., Bertelsen, M., Kucheryavsky, S. Predicting apple (cv. Elshof) postharvest dry matter and soluble solids content with near infrared spectroscopy. J Sci Food Agric. Jsfa6343 (2013)

[10] Liu F, Jiang Y, He Y, Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determinate soluble solids content of beer. Anal. Chim. Acta. 635: 45-52 (2009)

[11] Rossi F, Lendasse A, François D, Wertz V, Verleysen M, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. Chemom. Intell. Lab. Syst. 80: 215-226 (2006)

[12] Pent H, Long F, Ding C, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Analysis and Mach Intell 27(8):1226-1238 (2005)

[13] Suykens JAK, Vanderwalle J, Least squares support vector machine classifiers. Neural process. Lett. 9:293-300 (1999)

[14] Cover TM, Thomas JA. Elem. Inform. Theory. Wiley, New York (1991)

[15] Estévez PA, Tesmer A, Perez CA, Zurada JM, Normalized mutual information feature selection, IEEE Trans. Neural Netw 20: 189–201 (2009)

[16] Guyon I, Elisseeff A, An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3: 1157-1182 (2003)

[17] Koller D, Sahami M, Toward Optimal Feature Selection. Proc. Int. Conf. on Machine Learning, 284-292 (1996)

[18] Herrera LJ, Pomares H, Rojas I, Verleysen M, Guillén A, Effective Input Variable Selection for Function Approximation. Lecture notes in computer science, Springer, ICANN P, 4131:41-50 (2006)

[19] Kraskov A, Stgbauer H, Grassberger P, Estimating mutual information. Phys.Rev.,E 69, 066138 (2004)

[20] Moons E, Dardenne P, Dubois A, Sindic M, Non-destructive visible and NIR spectroscopy measurement for the determination of apple internal quality. Acta Hort 517: 441-448 (1997)

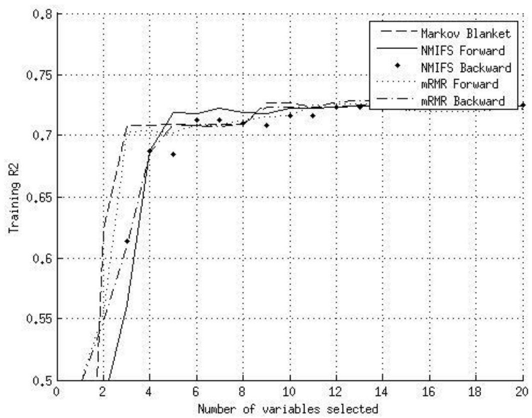
[21] Tu K, De Busscher R, De Baerdemaeker J, Schrevens E, Using laser beam as light source to study tomato and apple quality non-destructively. Food Processing Automation IV Conferenc. Chicago, IL, 3-5: 528-536 (1995)

[22] Damien François, Fabrice Rossi, Vincent Wertz, Michel Verleysen, Resampling methods for parameter-free and robust feature selection with mutual information, Neurocomputing 70, 7-9 (2007) 1276-1288

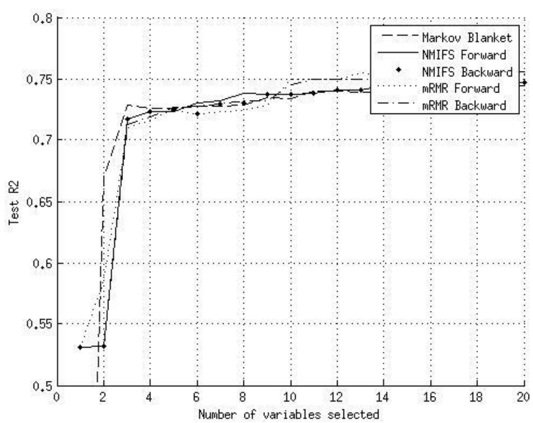
[23] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 42(2), (2012) 513-529

[24] Victoria Lafuente, Luis Javier Herrera, María del Mar Pérez, Jesús Val, Ignacio Negueruela: Firmness prediction in Prunus persica 'Calrico' peaches by Visible/short wave near infrared spectroscopy and acoustic measurements using optimized linear and non-linear chemometric models, Journal of the Science of Food and Agriculture, doi: 10.1002/jsfa.6916. 2014

[25] Ruiz-Altisent M, Lleo L and Riquelme F, Instrumental quality assessment of peaches: Fusion of optical and mechanical parameters. J Food Eng 74:490–499 (2006).

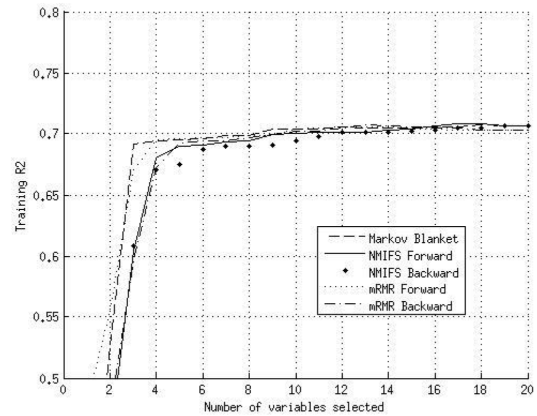


a)

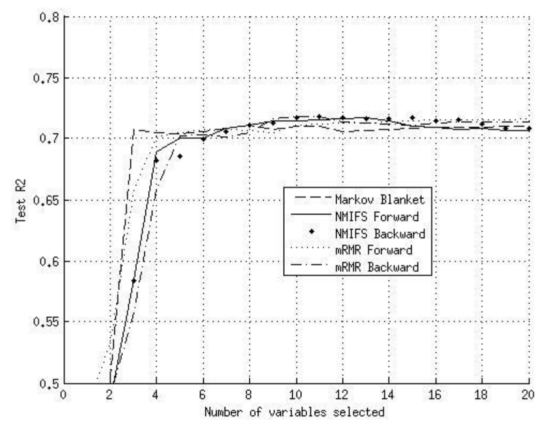


b)

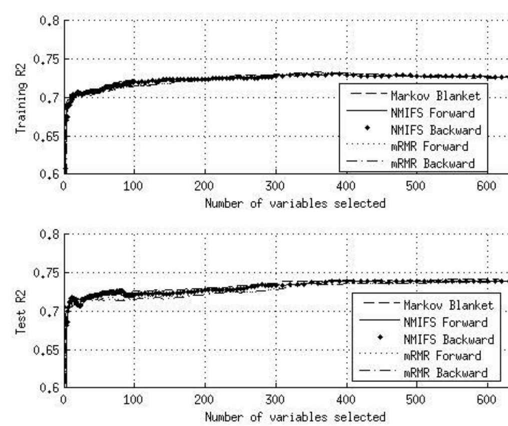
Fig. 1. Feature selection performance in the second problem of the five algorithms considered in this work (Markov Blanket, NMIFS Forward, NMIFS Backward, mRMR Forward and mRMR Backward) for one of the 10-fold CV executions. Both for training CV error (**a**) and for test error (**b**), zoom is displayed for 1-20 possible subsets of variables.



a)



b)



c)

Fig. 2. Feature selection mean performance in the second problem of the five algorithms considered for the 10 executions in the 10-fold CV executions. Both for training CV error (**a**) and for test error (**b**). Zoom is displayed for 1-20 possible subsets of variables, and complete overview can be seen in subfigure (**c**).

TABLE II
5-FOLD TEST CROSS VALIDATION R2 VALUES FOR THE 7 PROBLEMS CONSIDERED FOR DIFFERENT NUMBER OF VARIABLES SELECTED

	# vars	Markov Blanket	Forward NMIFS	Backward NMIFS	Forward mRMR	Backward mRMR
Apple Firmness	1	0,64	**0,67**	0,61	**0,67**	0,60
	2	0,70	0,71	0,67	**0,72**	0,65
	5	0,74	**0,77**	0,74	**0,77**	0,73
	10	0,76	**0,79**	0,78	**0,79**	0,77
	50	0,81	**0,82**	0,81	**0,82**	0,81
	100	0,82	0,82	0,82	**0,83**	0,82
	All	0,83	0,83	0,83	0,83	0,83
Apple SSC	1	0,61	0,62	0,62	0,62	**0,63**
	2	**0,70**	0,62	0,63	0,62	0,63
	5	**0,77**	0,68	0,69	0,73	0,72
	10	**0,80**	0,73	0,73	0,79	0,78
	50	**0,86**	0,82	0,82	**0,86**	0,85
	100	**0,87**	0,85	0,84	**0,87**	**0,87**
	All	0,88	0,88	0,88	0,88	0,88
Calrico Firmness	1	0,23	0,23	**0,26**	0,23	**0,26**
	2	**0,45**	0,38	0,39	0,37	0,40
	5	**0,59**	0,50	0,51	0,54	0,52
	10	**0,64**	0,57	0,58	0,61	0,60
	50	**0,70**	0,67	0,67	0,69	0,69
	100	**0,71**	0,70	0,70	**0,71**	**0,71**
	All	0,74	0,74	0,74	0,74	0,74
Calrico Aweta	1	0,38	0,38	**0,39**	0,38	**0,39**
	2	**0,45**	0,42	0,43	0,42	0,43
	5	**0,51**	0,49	0,47	0,49	0,50
	10	**0,53**	0,52	0,50	0,52	**0,53**
	50	**0,57**	0,55	0,55	**0,57**	**0,57**
	100	**0,58**	0,56	0,56	**0,58**	**0,58**
	All	0,59	0,59	0,59	0,59	0,59
Cashmere SSC	1	0,30	**0,45**	0,43	**0,45**	0,43
	2	0,50	0,46	0,45	**0,54**	0,46
	5	0,65	0,56	0,54	**0,66**	0,63
	10	**0,71**	0,65	0,62	**0,71**	0,70
	50	0,78	0,77	0,76	**0,80**	0,79
	100	0,81	0,79	0,79	**0,82**	0,81
	All	0,85	0,85	0,85	0,85	0,85
Chelan SSC	1	0,40	**0,41**	0,37	**0,41**	0,39
	2	**0,52**	0,43	0,39	0,48	0,41
	5	**0,63**	0,55	0,53	0,61	0,55
	10	**0,68**	0,62	0,60	0,66	0,63
	50	**0,73**	0,71	0,70	0,71	0,69
	100	**0,75**	0,73	0,72	0,71	0,70
	All	**0,80**	0,76	0,76	0,76	0,76

**Revised manuscript April 2015. Changes: equations in page 2 and 3 fixed.