USE OF FEATURE TRANSFORMATION AND FEATURE SELECTION PROCESSES BEFORE MODELLING VARIABLES RELATED WITH VITICULTURE CROPS

Roberto Fernández Martínez

Julio Fernández Ceniceros

Andrés Sanz García

Marina Corral Bobadilla

Universidad de La Rioja

Abstract

Nowadays, monitoring methods carried out in agricultural and environmental processes can obtain a large amount of data from different variables. The features improvements in devices that are responsible for measuring have allowed the possibility of collecting data from a large number of features and with better accuracy. This means that it is had a huge amount of variables and data stored from which it is possible to work with. Therefore applying learning algorithms becomes more complicated as the amount of data stored is higher, increasing the execution time and errors produced by the algorithms used. To solve this problem it is worked on different methods to reduce the number of input variables to use in these algorithms. This kind of methods can be divided mainly into two categories: feature transformation and feature selection. Using these algorithms, the variables are transformed or selected to improve the results of the subsequent research. Specifically, in this study, these processes are applied to environmental variables that affect a crop during its ripening, as a vineyard.

Keywords*: feature transformation; feature selection; environmental monitoring; viticulture crops.*

Resumen

En la actualidad, las monitorizaciones realizadas en procesos agrícolas y medioambientales permiten captar una gran cantidad de datos de diferentes variables. La mejora de las características de los aparatos responsables de estas mediciones ha permitido la posibilidad de obtener datos de una mayor cantidad de características y con una mayor precisión. Esto implica que se dispone de una gran cantidad de variables y de datos almacenados con los que trabajar. Por tanto, aplicar algoritmos de aprendizaje se vuelve más complicado conforme la cantidad de datos almacenada es mayor, aumentando los tiempos de ejecución y empeorando los errores de los algoritmos utilizados. Para solucionar este problema se está trabajando en diferentes métodos que reduzcan las variables de entrada de los algoritmos. Estos métodos pueden diferenciarse principalmente en dos categorías*: feature transformation y feature selection*. A través de estos algoritmos las variables son transformadas o seleccionadas para mejorar los resultados del trabajo a realizar posteriormente. Concretamente, dentro de este estudio, estos procesos son aplicados a

variables medioambientales que influyen en la maduración de un cultivo de temporada, como es un viñedo.

Palabras clave: *Transformación y selección de variables; monitorización medioambiental; cultivos vitivinícolas*

1. Introducción

Actualmente las nuevas tecnologías permiten recoger multitud de datos a través de sensores que recogen variables físicas y monitorizan cultivos agrícolas durante su época de crecimiento y maduración (Hwang et al., 2010; Mazzetto at al., 2010). A partir de estos datos, y analizando todas estas variables recogidas se puede obtener un conocimiento que ayude en la mejora del proceso y en la toma de decisiones de su gestión (Cardenas et al., 2010).

Los análisis de los datos recogidos se realizan usando técnicas de minería de datos que permiten convertir esos datos en información útil que ayude al control del cultivo (Ceglar et al., 2010; Zang et al., 2003; Stockle et al., 2003). Pero cuando la cantidad de datos recogida es muy elevada, aparece el problema de que muchos de ellos son irrelevantes e incluso perjudiciales a la hora de aplicar algoritmos de aprendizaje y trabajar con ellos.

Ante estos problemas, previamente a la utilización de algoritmos de aprendizaje, se trabaja con herramientas de transformación y selección de variables que permite reducir el número de variables eliminando las que tienen información irrelevante o redundante (Piramuthu, 2004).

En el caso de monitorización agrarias son muchas las variables meteorológicas que son recogidas y que influyen en le proceso. En este trabajo se realiza un estudio para saber cuales de las variables van a afectar de manera más significativa al cultivo, para de esta manera poder realizar una reducción de variables que permita mejorar la calidad en la posterior aplicación de algoritmos de aprendizaje.

En el caso de estudio se determina cuales de las variables meteorológicas relacionadas con varios viñedos de la Denominación de Origen Calificada Rioja afectan a atributos obtenidos durante la maduración de las uvas. Se realiza un estudio analizando como estas variables afectan a la acidez de las uvas, aunque utilizando previamente técnicas de transformación y selección de variables. Lo cual permite reducir el número de variables con el que poder estudiar el parámetro de acidez de las uvas del viñedo y mejorar la calidad de este futuro trabajo. Una vez se seleccionan las variables se realiza un análisis de cómo afectará esta selección a un posible análisis de la acidez.

2. Área de estudio

Los datos utilizados en este estudio, han sido recogidos de diferentes zonas de estudio de La Rioja (España). Una vez recogidos, se han generado varias de las variables consideradas necesarias para controlar este proceso, según recomiendan varios autores (Ribèreau-Gayon et al, 2006; Jackson, 2008; Coombe, 1992). Llegando a la conclusión de que las variables útiles para predecir estas características son 29. Disponiendo además, de cada una de estas variables, de una gran cantidad de información ya que el estudio se realiza con datos recogidos durante 7 años.

Las variables sobre las que se realiza el estudio son las mostradas en la Tabla 1.

Tabla 1: Variables seleccionadas de partida para el entrenamiento de modelos.

Vineyard variables	
Location	Loc
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

Para la comparación de los resultados se ha utilizado una red neuronal, la cual permite entrenar modelos basados en ésta con las variables seleccionadas. Usando las variables utilizadas desde el inicio los estimadores de error obtenidos se muestran en la Tabla 2.

3. Transformación de atributos

Dentro de los procesos de transformación de atributos se engloban todos los que modifican la forma de los datos. En el caso de estudio se realiza una transformación de datos con objeto de reducir la dimensionalidad de los datos de partida, ya que una alta dimensionalidad puede suponer un problema para una posterior aplicación de algoritmos de aprendizaje sobre los datos.

Tabla 2: Errores obtenidos utilizando todas las variables seleccionadas inicialmente

	Tiempo	Error testeado con red neuronal				
		CORR	MAE	RMSE	RAE	RRSE
Todos los datos	47.77	0.8432	0.0524	0.0756	59.1588	59.2657

Para solucionarlo se realiza una transformación de los datos, que llamaremos proyección, y que sustituye el conjunto inicial de datos por otro conjunto de datos transformados. Existen multitud de técnicas que resultan útiles para este proceso (Carreira-Perpiñán, 1997):

- análisis de componentes principales (PCA)
- low dimensional projection of the data (projection pursuit, generalised models)
- regression (principal curves)
- self-organization (Kohonen's maps)
- topology continuous mappings (generative topographic mapping)

3.1 Análisis de componentes principales (PCA)

En este trabajo se utiliza el método PCA (Jolliffe, 2002) ya que de todos los relacionados con la transformación de atributos es el más eficiente. Este método consiste en transformar las n variables originales A_1, A_2, A_3,… A_n en otro conjunto de atributos B_1, B_2, B_3,…,B_m donde m<n eliminando la menor cantidad de información posible.

Lo interesante de este método es que los nuevos elementos se han generado de manera que son independientes entre si. Estas nuevas variables no están correladas entre si por lo que no ofrecen redundancia en la información contenida. Además se ordenan por orden de relevancia, por lo que se pueden elegir los k primeros atributos que más relevancia tengan y de esa manera reducir el número de variables de entrada en trabajos posteriores. Consiguiendo que la varianza, o variabilidad de los datos, de los nuevos atributos sea mayor que la de los iniciales.

Posteriormente, se realiza un análisis comparando cómo la eliminación de variables menos significativas afecta a un posterior trabajo con los datos. Se van eliminando estas variables y se analizan los datos obteniendo un error y un tiempo de ejecución. De esta manera podemos comprobar cuando se obtiene la mejor de las configuraciones posibles.

A la hora de aplicar los algoritmos de transformación se separan las variables en grupos de variables (Tabla 3) relacionadas entre si para conseguir unos mejores resultados, separando varias variables de las iniciales, las cuales no se transformarán, pues se consideran importantes.

Tabla 3: Grupos en que se dividen las variables para utilizar PCA

Grupos utilizados en PCA	Variables que forman parte de cada grupo
G1 - LLuvAcuSem	RFW, RF2W, RF3W, RFW2, RF2W2
G2 - Temp	Temp, TempMin, TempMax, Stemp, D40, DM18, DM30, DDN
G3 - LLuvAcuPeriodo	RFY, RFBB, RFBBF, RFFS, RFSV, RFVH
G4 - Resto	Dir, Hum, HumMin, HumMax, Speed, SpeedMax

Los resultados obtenidos al aplicar el algoritmo PCA son los mostrados en la Tabla 4.

Una vez transformadas las variables se aplica un algoritmo basado en redes neuronales para comprobar cual es el comportamiento de la transformación. La evaluación de la selección desarrollada se realizo estudiando los siguientes estimadores: CORR (Correlation), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), RAE (Relative Absolute Error), RRSE (Root Relative Squared Error). La correlación y los errores obtenidos se muestran en la Tabla 5.

Tabla 4: Resultados obtenidos al aplicar PCA, ordenados por la proporción de varianza que explican.

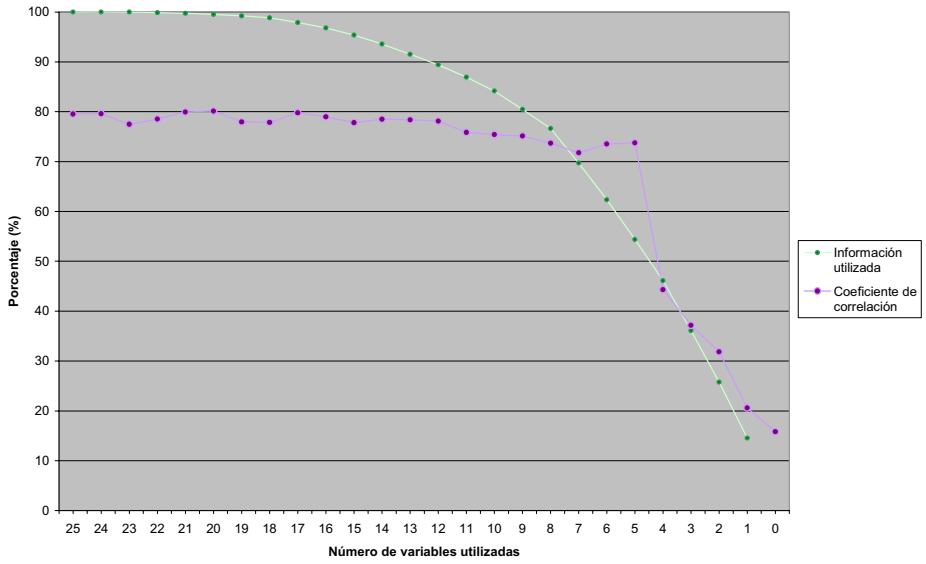
	Proporción de varianza explicada por grupos	Proporción acumulada por grupos	Proporción de varianza total	Proporción acumulada total
LluvAcuSemPC5	0.000	0.000	0.000	0.000
LluvAcuSemPC4	0.000	0.000	0.000	0.000
TempPC8	0.004	0.004	0.001	0.001
LLuvAcuPeriodoPC6	0.006	0.010	0.001	0.003
RestoPC6	0.009	0.018	0.002	0.005
TempPC7	0.010	0.028	0.003	0.008
LLuvAcuPeriodoPC5	0.016	0.044	0.004	0.012
RestoPC5	0.039	0.083	0.009	0.021
TempPC6	0.033	0.116	0.011	0.032
LLuvAcuPeriodoPC4	0.061	0.177	0.015	0.046
LluvAcuSemPC3	0.089	0.266	0.018	0.064
RestoPC4	0.086	0.352	0.021	0.085
TempPC5	0.066	0.418	0.021	0.106
RestoPC3	0.104	0.522	0.025	0.131
TempPC4	0.086	0.608	0.028	0.158
LluvAcuSemPC2	0.184	0.792	0.037	0.195
LLuvAcuPeriodoPC3	0.161	0.953	0.039	0.234
LLuvAcuPeriodoPC2	0.288	1.241	0.069	0.303
TempPC3	0.230	1.471	0.074	0.376
RestoPC2	0.332	1.803	0.080	0.456
TempPC2	0.259	2.062	0.083	0.539
TempPC1	0.312	2.374	0.100	0.639
RestoPC1	0.431	2.805	0.103	0.742
LLuvAcuPeriodoPC1	0.468	3.273	0.112	0.855
LluvAcuSemPC1	0.727	4.000	0.145	1.000

Tabla 5: Errores producidos en el entrenamiento de modelos dependientes del número de atributos seleccionados. Ordenados según la información recogida.

Número de atributos	Información recogida (%)	Tiempo de ejecución	Error testeado con red neuronal				
			CORR	MAE	RMSE	RAE	RRSE
n	100.000	60.58	0.7953	0.0539	0.0783	61.22	61.78
n-1	99.999	60.20	0.7960	0.0546	0.0784	62.10	61.87
n-2	99.999	53.38	0.7749	0.0592	0.0830	67.34	65.44
n-3	99.871	51.36	0.7854	0.0559	0.0805	63.48	63.49
n-4	99.733	73.05	0.7996	0.0544	0.0777	61.88	61.29
n-5	99.527	50.66	0.8014	0.0537	0.0766	61.02	60.41
n-6	99.204	42.61	0.7796	0.0563	0.0818	63.94	64.56
n-7	98.820	51.48	0.7787	0.0558	0.0810	63.42	63.87
n-8	97.891	36.42	0.7979	0.0532	0.0771	60.49	60.80
n-9	96.832	31.75	0.7898	0.0563	0.0793	63.95	62.53
n-10	95.368	30.17	0.7782	0.0585	0.0813	66.46	64.16
n-11	93.588	30.30	0.7850	0.0570	0.0800	64.76	63.07
n-12	91.524	25.99	0.7838	0.0553	0.0794	62.84	62.60
n-13	89.422	22.63	0.7811	0.0566	0.0807	64.33	63.65
n-14	86.926	20.88	0.7585	0.0612	0.0847	69.57	66.84
n-15	84.174	18.92	0.7542	0.0619	0.0869	70.42	68.52
n-16	80.494	16.91	0.7514	0.0603	0.0857	68.50	67.61
n-17	76.630	14.45	0.7370	0.0640	0.0886	72.74	69.86
n-18	69.718	15.03	0.7177	0.0676	0.0913	76.82	72.04
n-19	62.358	11.25	0.7352	0.0648	0.0892	73.63	70.37
n-20	54.390	8.13	0.7375	0.0617	0.0866	70.09	68.31
n-21	46.102	5.84	0.4432	0.0848	0.1179	96.35	93.03
n-22	36.118	5.91	0.3718	0.0902	0.1241	102.48	97.89
n-23	25.774	5.79	0.3185	0.0900	0.1255	102.34	98.99
n-24	14.542	5.80	0.2059	0.0911	0.1301	103.52	102.65
Resto			0.1582	0.0916	0.1319	104.11	104.03

Los errores no varían demasiado, pero la utilización de este método determina que podemos usan menos variables para obtener errores muy parecidos (Figura 1). Y a igualdad de precisión es preferible usar un modelo con menos características ya que permiten evitar problemas en el entrenamiento de futuros modelos y mejorar los rendimientos en cuanto a tiempo.

Figura 1: Visualización de la correlación final obtenida respecto a la reducción de variables transformadas con algoritmos PCA.



4. Selección de atributos

Estos métodos consisten en buscar un subconjunto de atributos originales que mejoren la calidad del modelo y permitan describir de manera satisfactoria el conjunto total de atributos (Liu and Motoda, 2008). La gran ventaja con respecto a la transformación de atributos, es que en éste las variables que forman el modelo son variables originales sin transformar.

Los métodos de selección de atributos muestran varias ventajas características:

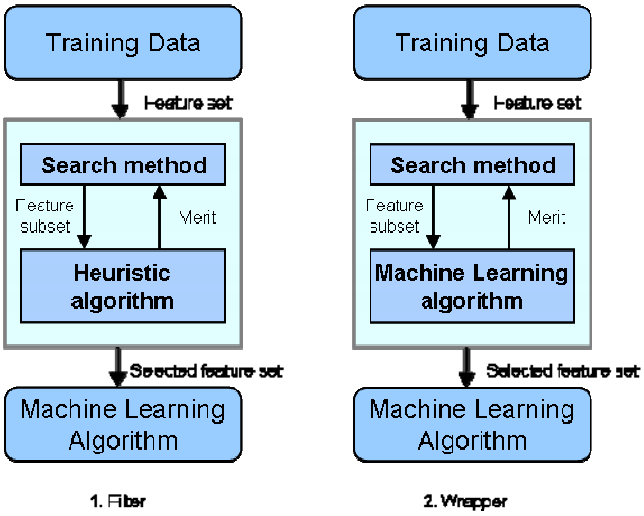
- Permiten reducir el tamaño de los datos eliminando atributos irrelevantes (su conocimiento no aporta nada al conjunto original de variables) o redundantes (puede ser determinada a partir de otras variables predictivas).
- Mejoran la calidad del modelo ya que los algoritmos que se utilizaran posteriormente solamente trabajan con las variables más influyentes.
- Se reduce el coste computacional en la aplicación posterior de algoritmos de aprendizaje.
- Al no realizarse una transformación de datos la claridad para comprender el modelo obtenido es mayor.
- Existe la posibilidad de reducir las variables hasta poder realizar una visualización de los datos.

Se eliminan campos donde el dato es una clave primaria o única para cada registro, y también campos donde los atributos sean dependientes.

Existen dos grupos de métodos para la selección de variables (Guyon and Elisseeff, 2003).

- Filters o aproximación indirecta: En este método se filtran los atributos irrelevantes antes de realizar cualquier proceso de minería de datos, usando técnicas fundamentalmente estadísticas (Duch, 2006). Se establece un ranking, según la técnica seleccionada, entre las variables disponibles según la relevancia que tienen sobre el conjunto total de datos. Se utilizan algoritmos heurísticos, los cuales no determinan la solución de forma directa, sino mediante ensayos y pruebas.
- Wrappers o aproximación directa: la selección se realiza según un modelo de minería de datos. Por lo que utilizan más tiempo ya que se requiere entrenar un modelo (Kohavi and John, 1998). Se selecciona mediante una búsqueda, un subconjunto de elementos además de la variable determinante, para estudiar el subconjunto empleando algoritmos de aprendizaje.

Figura 2: Diagrama de bloques explicativo del funcionamiento de los métodos filter y wrapper.



Ambos métodos, tanto filters como wrappers, necesitan métodos de búsqueda para seleccionar los atributos con los que trabajar, seleccionando un subconjunto de m parámetros de entre un conjunto original de n parámetros candidatos, con m<n. Ambos métodos son iterativos, lo que quiere decir que se van eligiendo atributos y se va observando el resultado, hasta que se obtiene la combinación de variables que mejores resultados ofrece.

Existen dos estrategias básicas para realizar la búsqueda (Guodong and Ganlin, 1988):

- Estrategias forward: Se empieza con el atributo que de mejores resultados según el algoritmo utilizado por el método de búsqueda. Por ejemplo, en una regresión lineal se selecciona la característica más correlacionada con la variable de salida. Se van añadiendo atributos de la misma manera, seleccionando el que mejor resultado ofrezca unido a los ya seleccionados. Repitiendo el proceso hasta que se obtiene la mejor calidad total (Atkinson et al., 2010).
- Estrategias backward: Se comienza con el conjunto de datos completo para en cada paso ir eliminando una característica. Se realizan selecciones aleatorias de atributos eliminando una característica hasta que se encuentra la que ofrece mejores resultados.

La variable eliminada en esta configuración será eliminada del subconjunto de datos y ya no volverá a formar parte de él.

En el caso práctico que nos ocupa se han aplicado el filtro CFS de tipo SubSet, evaluador de atributos, con dos métodos de búsqueda diferentes, best first y genetic search. Y el método ReliefF de tipo AttributeEval, evaluador de atributos, con métodos de búsqueda tipo ranker, los cuales hacen búsqueda de atributos simples ordenándolos por relevancia. Los métodos wrappers han sido descartados puesto que los tiempos de ejecución de estos eran demasiado elevados.

4.1 Correlation-based Feature Subset Selection (CFS)

Este tipo de filtro (Hall, 1998) evalúa el valor de un subconjunto de los atributos, considerando la capacidad individual de predicción de cada una de las características teniendo en cuenta el grado de redundancia existente entre ellos.

Son elegidos los subconjuntos de características que están altamente correlacionadas con la clase y que tengan una baja intercorrelación entre ellos.

Según la formula:

$$G_s = \frac{k \cdot \bar{r}_{ci}}{\sqrt{k \cdot (k-1) \cdot \bar{r}_{ii}}} \quad (1)$$

Donde k es el número de atributos del subconjunto, $\bar{r}_{ci}$ es la correlación media con la clase y $\bar{r}_{ii}$ es la correlación media de los atributos entre si. Puede considerarse que el numerador define el nivel de predicción de la clase con el subconjunto seleccionado y el denominador la redundancia existente entre atributos (Hall y Smith, 1999).

Para el estudio realizado se ha utilizado este método con dos algoritmos de búsqueda, Best First y Genetic Search. Mostrando los resultados en la Tabla 6.

Tabla 6: Resultados de selección de variables obtenidos con el filtro de tipo Cfs.

Cfs Subset Evaluation			
Best First		Genetic Search	
Porcentaje	Atributos	Porcentaje	Atributos
100	Var	100	Var
100	Altit	100	Altit
100	TempMax	100	RFBBF
100	DM18	100	DM18
100	DM30	100	DM30
100	RFBBF	90	RFSV
90	RFSV	80	TempMax
10	HumMax	20	HumMax
		10	RF2W
		10	RFW2
		10	Temp
		10	TempMin

Una vez entrenado el modelo basado en redes neuronales, con las variables seleccionadas, se obtienen los errores mostrados en la Tabla 7.

Tabla 7: Errores producidos en el entrenamiento de modelos dependientes del número de atributos seleccionados con el filtro de tipo Cfs.

Algoritmo	Método de busqueda	Tiempo	Error testeado con red neuronal				
			CORR	MAE	RMSE	RAE	RRSE
Cfs Subset Evaluation	Best First	9.05	0.7352	0.0746	0.0998	84.1316	78.1939
Cfs Subset Evaluation	Genetic Search	14.02	0.7558	0.0708	0.0956	79.9196	74.9428

4.2 ReliefF

El método ReliefF (Kira y Rendell, 1992) realiza la selección de atributos de acuerdo a cómo selecciona los vecinos más cercanos de la misma clase comparándolo con cómo selecciona los vecinos más cercanos de otras clases diferentes. Se realiza una búsqueda de n vecinos más cercanos, siendo uno de ellos de la misma clase y los demás de clases diferentes. A partir de estos valores se estima el valor W(A) según la diferencia de probabilidades (Kononenko, 1994):

$$W(A) = P(valor\ diferente\ de\ A \mid C_X = C_A) - P(valor\ diferente\ de\ A \mid C_X \neq C_A) \quad (2)$$

La Ecuación 2 representa el algoritmo Relief que posteriormente ha sido extendido, pasando a denominarse ReliefF, donde se aplica el mismo principio, pero analizando todas las clases restantes.

$$W(A) = W(A) - \frac{diff(A,R,H) + \sum_{C \neq Clase(R)} P(C) \cdot diff(A,R,M(C))}{l} \quad (3)$$

Donde $diff(A,R,H)$ es la distancia entre el atributo (A) y las instancias consideradas de la misma clase (R y H) y $diff(A,R,M(C))$ es la distancia entre el atributo y las instancias más cercanas de otras clases.

Para el estudio realizado se ha utilizado este método con el algoritmo de búsqueda tipo ranker, mostrando los resultados en la Tabla 8.

Tabla 8: Resultados de selección de variables obtenidos con el filtro de tipo ReliefF.

ReliefFAttributeEval	
Ranking	Atributos
0.1063186	Var
0.0118399	W100B
0.0076087	Altit
0.0022264	RFSV
0.0010835	DM18
0.0008251	D40
0.0003259	RFBB
0.0000126	RFY

Una vez entrenado el modelo basado en redes neuronales, con las variables seleccionadas, se obtienen los errores mostrados en la Tabla 9.

Tabla 9: Errores producidos en el entrenamiento de modelos dependientes del número de atributos seleccionados con el filtro de tipo ReliefF.

Algoritmo	Metodo de busqueda	Tiempo	Error testeado con red neuronal				
			CORR	MAE	RMSE	RAE	RRSE
ReliefFAttributeEval	Ranker	9.07	0.7909	0.0637	0.0871	71.8737	68.2615

5. Conclusiones

Se observa que la utilización de métodos de transformación y selección de variables ha sido efectiva ya que al reducir las variables de entrada se ha mejorado la comprensión de los modelos generados. Y puesto que el modelo está compuesto de menos variables, son las más significativas las que definen el problema de una manera más clara.

También se ha comprobado una reducción del gasto computacional a la hora de generar el modelo final. Esto sucede sobre todo en los modelos más complejos, puesto que no es lo mismo calibrar modelos con muchas variables que no aporten información significativa, que calibrar sólo con las más significativas. Los tiempos han sufrido una reducción considerable al reducir la carga de entrada a los algoritmos de aprendizaje.

Como conclusión final se ha determinado que la aplicación de estos algoritmos ha resultado útil con esta clase de datos y se aconseja para futuros trabajos.

6. Referencias

Atkinson, A.C., Riani, M., Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society*, Vol. 39 (2), pp. 117-134.

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

Carreira-Perpiñán, M. A. (1997). A Review of Dimension Reduction Techniques. *Technical report CS-96'09*. Dept. of Computer Science. University of Sheffield.

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

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

Duch, W. (2006). Chapter 3: Filter Methods. *Feature extraction, foundations and applications*. Springer. Pp. 89-118.

Guodong, Z., Ganlin, Y. (1988). Forward-backward search method. *Journal of Computer Science and Technology*, Vol. 3, Number 4, pp. 289-305.

Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, Vol. 3, pp. 1157-1182.

Hall, M. A. (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

Hall, M. A., Smith, L.A. (1999). Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, Orlando, USA, pp. 235-239.

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

Jackson, R.S. (2008). *Wine Science Principles and Applications*, Third Edition. Elsevier Inc.

Jolliffe, I.T. (2002*). Principal Component Analysis*, Series: Springer Series in Statistics. 2nd ed. Springer, NY.

Kira, K., Rendell, L.A., (1992). A Practical Approach to Feature Selection. D. H. Sleeman & P. Edwards, eds, *Ninth International Workshop on Machine Learning*. Morgan Kaufmann, pp. 249-256.

Kohavi, R., John, G. (1998). The Wrapper Approach. *Feature Extraction, Construction and Selection: A Data Mining Perspective*, edited by Huan Liu and Hiroshi Motoda.

Kononenko, I., (1994). Estimating Attributes: Analysis and Extensions of RELIEF. F. Bergadano & L. De Raedt, eds, *European Conference on Machine Learning*. Springer pp. 171-182.

Liu, H., Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

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

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

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

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

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

Agradecimientos

Los autores agradecen a la “Dirección General de Investigación” del Ministerio Español de Ciencia e Innovación por el apoyo financiero de los proyectos DPI2007-61090; y a la Unión Europea por el proyecto RFSPR-06035.

Finalmente, los autores quieren agradecer también al Gobierno Autonómico de La Rioja por su apoyo a través del 3º Plan Riojano de I+D+i por el proyecto FOMENTA 2010/13 y a la Universidad de La Rioja por su ayuda a través de sus becas FPI.

Correspondencia (Para más información contacte con):

Roberto Fernández Martínez.
Área de Proyectos de Ingeniería. Departamento de Ingeniería Mecánica.
Universidad de La Rioja
C/ Luis de Ulloa 20, 26004 Logroño, La Rioja (España).
Phone: +34 941 299 274
E-mail: roberto.fernandez@unirioja.es
URL: http://www.mineriadatos.com