International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:4, 2011

Assessing and Visualizing the Stability of Feature Selectors: A Case Study with Spectral Data

R.Guzmán-Martínez, Oscar García-Olalla, and R. Alaiz-Rodríguez

Abstract—Feature selection plays an important role in applications with high dimensional data. The assessment of the stability of feature selection/ranking algorithms becomes an important issue when the dataset is small and the aim is to gain insight into the underlying process by analyzing the most relevant features. In this work, we propose a graphical approach that enables to analyze the similarity between feature ranking techniques as well as their individual stability. Moreover, it works with whatever stability metric (Canberra distance, Spearman's rank correlation coefficient, Kuncheva's stability index,...). We illustrate this visualization technique evaluating the stability of several feature selection techniques on a spectral binary dataset. Experimental results with a neural-based classifier show that stability and ranking quality may not be linked together and both issues have to be studied jointly in order to offer answers to the domain experts.

Keywords—Feature Selection Stability, Spectral data, data visualization.

I. Introduction

FEATURE selection is a key stage when working with multidimensional data [1]. In particular, spectral data are usually characterized by thousands of features whereas the data sets have a small number of instances due to the cost of recollection and labeling. Feature ranking or selection in order to reduce the data dimensionality is important in these applications for several reasons: (a) The curse of the dimensionality problem, since the size of the training data set needed to calibrate a model grows exponentially with the number of dimensions and (b) the knowledge extraction from the data is simplified if the instances are represented with less features. In particular, when working with spectral data, it is important to gain insight into the regions of the spectrum that have more discriminant power.

Feature selection or ranking techniques measure the importance of each feature according to the value of a given function. These algorithms can be divided in three types [2]: filter, wrapper and embedded approaches. The filter methods select the features according to a reasonable criterion computed directly from the data and that is independent of the classification machine. The wrapper approaches use the classification algorithm to determine the value of a given feature subset and the embedded techniques are specific for each model since they are intrinsically defined in the inductive algorithm.

R.Guzmán-Martínez is with the Servicio de Informatica y Comunicaciones. University of León. Campus de Vegazana s/n, 24071 León, Spain.

Oscar García-Olalla and R. Alaiz-Rodríguez are with the Dpto. de Ingeniería Eléctrica y de Sistemas. University of León. León, Spain.

Supported by the Spanish MEC project DPI2009-08424

A key issue of recent interest, is the stability of the feature selection and ranking algorithms. Fields like biomedicine or chemometrics, require not only accurate classification models, but a subset of the most important features in order to better understand the data and the underlying process. The fact that under small variations in the available training data, the top-k feature list varies, make this task not straightforward. To study the stability of the feature ranking/selector algorithms several (scalar) metrics have been proposed. The Spearman's rank correlation coefficient [3], [4] and Canberra distance [5] measure the similarity between rankings. When the goal is to measure the similarity between top-k lists (feature subsets), different authors have used the following meaures: Jaccard distance [3], Tanimoto distance [3], Kuncheva's stability index [6], Relative Hamming distance [7], Consistency measures, Dice-sorense's index, Ochiai's index or Percentage of overlapping features [8].

In general, the way these works proceed is a follows: Given a set of rankings (subsets), pairwise similarities are computed and then, reduced to a single metric by averaging. These (scalar) metrics can be seen as projections to one dimensional space and its use only shows where the feature selector stands in relation to the stable and the random ranking algorithm. In this paper we want to illustrate and motivate the use of graphical methods as a simple alternative approach to evaluate the stability of feature ranking algorithms. We will show how the projection to two dimensions allow to evaluate the similarity between feature ranking algorithms as well as their stability. We illustrate our approach with a study of six feature ranking algorithms on a fat spectral data set.

The rest of the paper is organized as follows. Section II and Section III formulate the feature selection problem and the study of its stability, respectively. Section IV discusses the advantages of the visual approach to this problem and Section V illustrates a study on a fat spectral dataset using both the classical analysis and the graphical approach and finally, Section VI summarizes the main conclusions.

II. Feature selection and Ranking

Feature selection techniques usually generate a full ranking of features. These rankings, however, can be converted in top-k lists that contain the k most important features.

Consider a training dataset $\mathcal{D} = \{(\mathbf{x}_i, d_i), i = 1, \ldots, M\}$ consisting of M instances and a target d associated with each sample . Each instance $\mathbf{x}_i$ is a p-dimensional vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots x_{ip})$ where each component x_{ij} represents the value of a given feature f_j for that example i, that is, $f_j(\mathbf{x}_i) = x_{ij}$.

Consider now a feature ranking algorithm whose output is a ranking vector $\mathbf{r}$ with components

$$\mathbf{r} = (r_1, r_2, r_3, \ldots, r_p) \tag{1}$$

where $1 \leq r_i \leq p$ and 1 is considered the highest rank.

The output of a feature selection algorithm is represented by a vector

$$\mathbf{s} = (s_1, s_2, s_3, \ldots, s_p), s_i \in \{0, 1\} \tag{2}$$

where 1 indicates the presence of a feature and 0 the absence and $\sum_{i=1}^{p} s_i = k$ for a top-k feature list.

Converting a ranking output into a feature subset is easily conducted according to

$$s_i = \begin{cases} 1 & \text{if } r_i \leq k \\ 0 & \text{if } otherwise \end{cases}$$

III. Stability of Feature Selectors

A problem that arises in many practical problems where knowledge is to be extracted from the ranking of features or the top-k features is that small variations in the data set leads to different outcomes. In particular, it is common when dealing with high dimensional data and few samples. Next, we present some common similarity metrics to measure the *distance* between two full rankings or two top-k lists.

A. Similarity Measures

Let $\mathbf{r}$ and $\mathbf{r}'$ be the outcome of a feature ranking algorithm applied to two different subsamples of $\mathcal{D}$. The most widely used metric to measure the similarity between two ranking outputs is the Spearman's rank correlation coefficient (SR). The SR between two rankings $\mathbf{r}$ and $\mathbf{r}'$ is defined as

$$SR(\mathbf{r}, \mathbf{r}') = 1 - 6 \sum_{i=1}^{p} \frac{(r_i - r_i')^2}{p(p^2 - 1)} \tag{3}$$

where r_i is the rank of feature-i. SR takes its value in the interval $[-1, 1]$. A value equal to -1 indicates exactly inverse orders, while a value of zero means no correlation. It takes the value of 1 when the rankings are identical.

When the goal is to measure the similarity between two partial lists $\mathbf{s}$ and $\mathbf{s}'$ with k features, several metrics have been proposed: Jaccard distance, Kuncheva's stability index, Relative Hamming distance, Consistency measures, Dice-sorense's index, Ochiai's index, Percentage of overlapping features (for details see [8]). Without loss of generality, in this work we focus on Kuncheva's stability index (KI) to measure the similarity between feature subsets. The KI between two top-k lists $\mathbf{s}$ and $\mathbf{s}'$ is defined as

$$KI(\mathbf{s}, \mathbf{s}') = \frac{rp - k^2}{k(p - k)} \tag{4}$$

where p is the original whole number of features, r is the number of features that are present in both lists simultaneously and $\sum_{i=1}^{p} s_i = \sum_{i=1}^{p} s_i' = k$. The KI satisfies $-1 < KI \leq 1$.

B. The Stability for a Set of Rankings or Lists

When we have a set of outputs from a feature selection (or ranking) algorithm, $\mathcal{A} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_K\}$, with size K, the most common way to evaluate the stability of the set is to compute pairwise similarities and average the results, what leads to a single scalar value.

$$S(\mathcal{A}) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} S_M(r_i, r_j) \tag{5}$$

where S_M represents any similarity metric (Kuncheva's stability index or Spearman rank correlation coefficient, for example).

IV. Visualizing the Stability of Feature Selectors

The outcome of a feature ranking algorithm can be interpreted as a point in a high dimensional space (with p dimensions). The stability of a ranking feature selector is commonly measured as the dissimilarity or distance between different outcomes of the same feature selector on slightly different datasets. That is, stability is assessed computing pairwise similarities between points in that high dimensional space and averaging the results. In this case, the ranking data is turned into a single number (projected to one dimension) and the algorithms are compared on the basis of this scalar metric. This only allows to compare the feature selector with respect to a reference: the random ranking and the completely stable ranking.

Note that if we change from a projection to an space with one dimension into a space with two or more dimensions, we have a visual representation that allows to establish comparisons with respect to the random or ideal feature selector as well as comparisons of each feature selector to the others.

In order to study the stability with a visual-based approach, different alternatives could be used, depending on the amount of information available. Note that, even simple visualization approaches (histograms, scatter graphs, spider graphs) allow to plot the results in a convenient way to ease result interpretation. They have some limitations as the number of dimensions increases. In this case, a dimensionality reduction technique like MDS [9], that preserves as much as possible the original data structure, seems more convenient. They allow to project data from a high dimensional space to a 2D or 3D space while preserving the distance in the original high dimensional space.

V. Typical Empirical Example with Spectral data

Consider a standard empirical study where L feature ranking algorithms give rise to K rankings, each one. The results for each feature ranking algorithm can be organized in a Table with elements r_{ij} with $i = 1, \ldots, p$ and $j = 1, \ldots, K$ that represent the rank assigned for the feature-i in the run-j

In this section a typical experiment is conducted in order to assess the stability of six feature selectors based on a filter approach : χ^2, Information Gain (IG), Information Gain Ratio (GR), Relief and other two based on the parameter values of an independent classifier (Decision Rule 1R and SVM). [10].

Experimental results were carried out with omental fat samples collected from carcasses of suckling lambs [11]. The

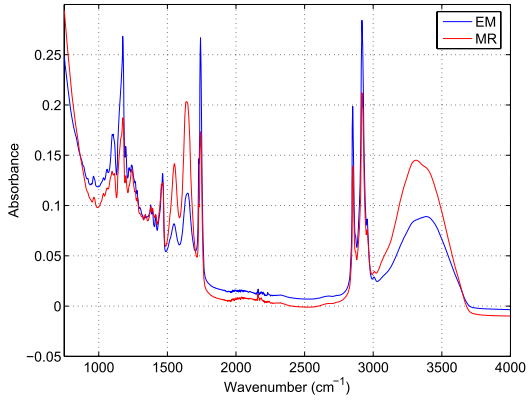

Fig. 1. Average FT-IR spectrum of omental fat samples for Milk Replacer (MR) and Ewe Milk (EM).

whole dataset has 134 instances: 66 from lambs being fed with a milk replacer (MR), while the other 68 are reared on ewe milk (EM). Authentication of the type of feeding will be a key issue in the certification of suckling lamb carcasses, with the rearing system being responsible for the difference in prices and quality. The use of spectroscopy for the discrimination of fat samples according to the rearing system provides several advantages, mainly its speed and versatility. Determining which regions of the spectrum have more discriminant power is also fundamental for the veterinarian professionals. All FTIR spectra were recorded from 4000 to 750 cm-1 with a resolution of 4 cm-1, what leads to a total of 1687 features. The average spectra for both classes is shown in Fig.1.

The dataset was randomly split in ten folds, launching the feature ranking algorithm with nine out the ten folds, in a consecutive way. Five runs of this process resulted in a total of $K = 50$ rankings. Feature ranking was carried out with WEKA [10] and the computation of the stability with MATLAB [12].

A. Traditional Stability Analysis

The stability of the feature ranking algorithms can be evaluated with metric like the Spearman's rank correlation coefficient (SR) in a conventional way. In this case, we have computed the $\frac{2}{50(50-1)}$ pairwise similarities for each algorithm to end up averaging these computations according to Eq.(5). The SR is recorded in Table I where it can be seen that Releif is the most stable (0.94) ranking algorithm, whereas 1R and SVM are quite unstable (0.79 and 0.74, respectively).

TABLE I
STABILITY OF A SET WITH 50 FULL RANKINGS ASSESSED THROUGH AVERAGE PAIRWISE SIMILARITIES WITH THE SPEARMAN'S RANK CORRELATION COEFFICIENT (SR).

1R	SVM	χ^2	GR	IG	Relief
0.79	0.74	0.86	0.90	0.86	0.94

The Kuncheva's stability index (KI) allows to study the stability of a feature subset that contains the top-k feature lists. Table II shows the KI for the selection of feature subsets with cardinality that varies from 10 to 1686.

TABLE II
STABILITY OF A SET WITH 50 TOP-K LISTS ASSESSED THROUGH AVERAGE PAIRWISE SIMILARITIES WITH THE KUNCHEVA'S STABILITY INDEX (KI) FOR DIFFERENT VALUES OF k.

k	1R	SVM	χ^2	GR	IG	Relief
10	0.301	0.026	0.785	0.785	0.785	0.746
50	0.633	0.097	0.829	0.829	0.829	0.676
100	0.743	0.175	0.839	0.840	0.837	0.691
200	0.852	0.303	0.912	0.845	0.910	0.771
300	0.765	0.401	0.847	0.799	0.839	0.842
400	0.693	0.461	0.776	0.768	0.782	0.844
500	0.656	0.510	0.758	0.801	0.774	0.841
600	0.631	0.543	0.737	0.812	0.758	0.823
700	0.609	0.565	0.715	0.790	0.744	0.820
800	0.590	0.587	0.681	0.783	0.733	0.786
900	0.575	0.600	0.655	0.786	0.720	0.753
1000	0.565	0.610	0.653	0.762	0.677	0.758
1100	0.553	0.614	0.674	0.702	0.643	0.797
1200	0.550	0.632	0.690	0.639	0.616	0.884
1300	0.523	0.646	0.641	0.614	0.563	0.930
1400	0.473	0.662	0.568	0.609	0.520	0.900
1500	0.388	0.674	0.518	0.563	0.512	0.887
1600	0.316	0.684	0.531	0.434	0.539	0.752
1686	0.542	0.959	0.456	0.473	0.340	0.725
Av.	0.577	0.513	0.698	0.718	0.691	0.801

The analysis based on a single metric does not allow, however, to say anything about how similar the rankings provided by the different algorithms are. Typical questions we would like to answer are: (ii) Which feature selector algorithms work similarly so that they can be considered equivalent?, (ii) Which feature algorithms provide very different rankings so that we have to evaluate their quality to induce a good performance classifier?, (iii) Which algorithm is more stable for a certain range of k values?. Analyzing directly the results gathered in Table II does not seem straightforward.

B. Visual-based Stability Analysis

1) Simple plots: A simple plot helps to see the relative and absolute stability of the feature selectors. Fig. 2 highlights that their relative stability changes with the value of k. In general terms, Releif appears to be the most stable algorithm. Note also that the stability of the SVM feature selector for low values of k is very low. No reliable information of the most important regions of the spectrum can be extracted from just a single run of the algorithm. It would be desirable to aggregate the rankings in order to get a more representative ranking. Likewise, 1R has also a high margin of stability improvement.

2) Assessment and visualization based on MDS: Multi-Dimensional Scaling (MDS) [9] is used in this section to visualize both the feature selectors as well as a completely random selector in a graph so that comparisons between all of them can be established.

All the results gathered in the experiment can be interpreted as a set of 300 points (6 algorithms x 50 runs each one) defined a 1687-dimensional space. These points are projected to a 2D space using MDS. The distance between points is calculated with the Spearman's rank coefficient and the stress criterion is normalized with the sum of squares of the dissimilarities.

After the projection, each outcome of the algorithm is represented by two coordinates (x,y) and the similarities among feature selector can be analyzed in Fig.3. In terms of

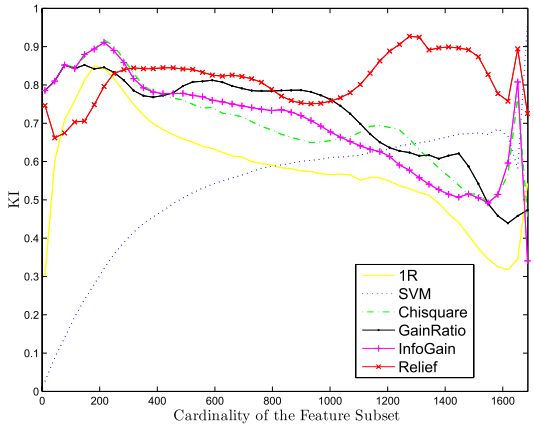

Fig. 2. KI for Feature Subsets with different cardinality

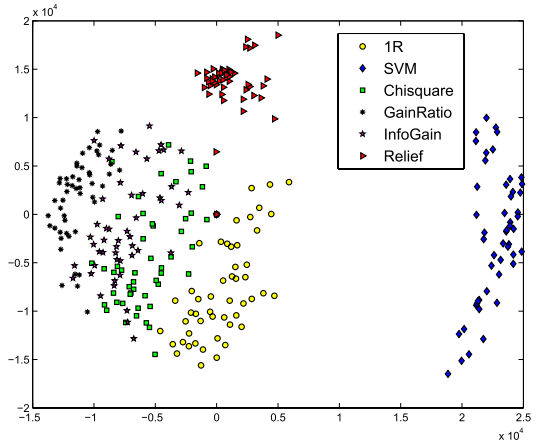

Fig. 3. MDS plot of the Feature Ranking Algorithms

stability we can see that the points that correspond to Releif are much less scattered than the rest. In other words, this is the most stable algorithm. The outcomes of SVM and 1R appear to be very scattered, what indicates their low stability. This graph allows to see that InfoGain generates similar ranking to GainRatio and χ^2 and that they can be considered equivalent in this context.

The ranking results can be also organized as a set of 6 points (6 algorithms) defined in a (1687 ranked features x 50 runs). An extra point corresponding to a random feature ranking algorithm can also be generated through simulation. Fig. 4 shows the distance to the Random selector. Releif is the most distant to a trivial random feature ranking algorithm. The figure also indicates that the ranking yielded by Releif is very different from the one generated by SVM and the other equivalent group (InfoGain, GainRatio, χ^2). This suggests that the rankings should be evaluated in order to determine the quality of the selected features to predict the target class (EM,MR). This is crucial in order to provide the veterinarian experts with reliable information about the most important regions of the spectrum, and not only with the most stable top-k list.

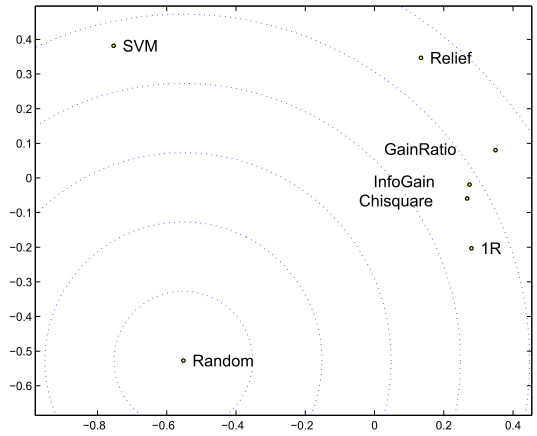

Fig. 4. MDS plot of the Feature Ranking Algorithms

C. Classifier Performance

A Multilayer Perceptron (MLP) has been used to classify the spectra. We evaluate a three layer network with a logistic sigmoid activation function for the hidden and the output layers. The (MSE, Mean Square Error) is the cost function minimized in the training stage. Several combinations of neurons in the hidden layer and different number of training cycles have been assessed with all the descriptors, in order to find the optimal network configuration (700 training cycles and 10 nodes in the hidden layer). The classifier error rate is estimated using 10-fold cross validation and the results shown in Table III are the average of 10 runs. Table III records the error of a classifier trained with the top-k features selected for the different ranking algorithms. The ranking used resulted from the aggregation of the 50 rankings by computing their median value.

TABLE III
ERROR RATE (IN %) FOR DIFFERENT FEATURE SELECTION ALGORITHMS.

Nmero of features	χ^2	GI	GIR	Relief	1R	SVM
20	16.64	19.48	17.69	15.92	17.3	11.11
40	14.46	13.54	16.81	15.26	16.08	10.62
80	14.08	13.75	19.42	16.12	14.61	10.55
160	12.56	13.11	15.67	17.6	12.40	9.43
300	12.64	12.81	17.18	17.21	13.42	**8.20**
600	13.11	13.18	13.87	12.42	10.86	9.07
900	13.46	13.58	12.71	10.45	**9.50**	8.71
1200	11.76	11.61	12.17	**9.54**	10.24	8.92
1500	**10.29**	**9.70**	**10.27**	10.00	9.54	9.46

As the previous analysis based on the MDS projections revealed, GI, GIR and χ^2 lead to similar performance since their rankings are very similar. Selecting the features according to Releif allows to get a classifier with lower error (9.54% with 1200 features). The ranking carried out with SVM, however, gets the lowest error, even though it is the least stable. Thus, the error is 8.20% with the top-300 features selected by the SVM feature selector algorithm.

Figure 5 plots the average spectrum of the omental fat dataset. Features are coloured according to their rank (median rank in the 50 runs): in black the regions with lowest rank

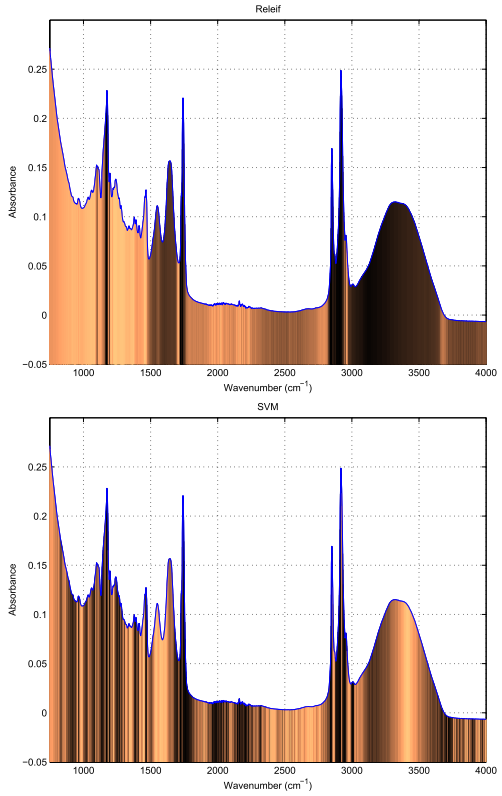

Fig. 5. Omental Fat Spectra. Ranking of the feature selctor algorithms Releif and SVM is shown. In black the regions with lowest rank and in light copper the ones with highest rank.

and in light copper the ones with highest rank. Even though the Releif algorithm is more stable, the ranking generated by SVM is more reliable to discriminate the fat samples. Feature ranking stability and its quality to induce good performance classifiers should be studied together for real practical applications. Information about the most discriminant regions of the spectra is useful to interpret what fat acids establish differences between animals reared by maternal milk and by milk replacers.

VI. Conclusion

In this work, we study the problem of robustness (or stability) of feature selection techniques. Based on several outcomes of a given feature ranking algorithm on slightly different data sets, traditional evaluation estimates its stability by computing an scalar metric. This can be viewed as a projection to a 1D space. We propose a graphical approach that works in 2D or 3D in order to evaluate not only the stability of the algorithms but also its similarities with other algorithms. Moreover, this enables to exploit the human visual capabilities in order to analyze the ranking or feature subsets.

We illustrate this technique on a fat spectra data set from suckling lambs (ones reared by maternal milk and others by milk replacers). This graphical approach based on a MDS projection allows to see at a glance and in a single picture that: (a) the most stable algorithm is Releif, (b) the most inestable is SVM, (c) the rankings yielded by IG, GR and χ^2 are very similar so that they can be considered equivalent. (d) The before mentioned group leads to a ranking that is very different to Releif and SVM.

Veterinarian experts are particularly interested in identifying the most discriminant regions of the spectra. Therefore, the predictive power of the top-k features using a neural network is evaluated as well. Experimental results show that stability and classifier performance are not linked together: in this case, the least stable algorithm leads to the most accurate classifier (on average). Future work includes the study of ensemble strategies to increase the stability of feature ranking algorithms, in particular those that have a high margin of improvement and evaluate the effect this has on classifier performance.

References

[1] I. Guyon, , and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=944968

[2] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[3] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, pp. 95–116, 2007.

[4] Y. Saeys, T. Abeel, and Y. Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," in *ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*. Springer-Verlag, 2008, pp. 313–325.

[5] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, "Algebraic stability indicators for ranked lists in molecular profiling," *Bioinformatics*, vol. 24, no. 2, p. 258, 2008.

[6] L. Kuncheva, "A stability index for feature selection," in *Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*. ACTA Press, 2007, pp. 390–395.

[7] K. Dunne, P. Cunningham, and F. Azuaje, "Solutions to instability problems with sequential wrapper-based approaches to feature selection," *Trinity College Dublin Computer Science Technical Report*, pp. 2002–28.

[8] Z. He and W. Yu, "Stable feature selection for biomarker discovery," Tech. Rep. arXiv:1001.0887, Jan 2010.

[9] T. Cox and M. Cox, *Multidimensional Scaling*. Chapman and Hall, October 1994.

[10] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, October 1999.

[11] M. Osorio, J. Zumalacrregui, R. Alaiz-Rodrguez, R. Guzman-Martnez, S. Engelsen, and J. Mateo, "Differentiation of perirenal and omental fat quality of suckling lambs according to the rearing system from fourier transforms mid-infrared spectra using partial least squares and artificial neural networks," *Meat Science*, vol. 83, no. 1, pp. 140 – 147, 2009.

[12] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

Roberto Guzman-Martinez received the BS degree in Industrial Engineering from the University of Burgos, Spain, in 2003 and he is currently working towards a Ph.D. degree from University of Leon, Spain. She is currently working in Servicio de Informatica y Comunicaciones, University of Leon, Spain. His research interests include statistical pattern recognition, data mining and its application to industrial applications.

Oscar García-Olalla received the BS degree in Computer Science from the University of León, Spain, in 2010 and he is currently a MSc. student at the Department of Electrical and Systems Engineering, University of León, Spain. His research interests include pattern recognition and image processing.

Rocío Alaiz-Rodríguez received the BS degree in Electrical Engineering from the University of Valladolid, Spain, in 1999 and the Ph.D. degree from Carlos III University of Madrid, Spain. She is currently an Associate Professor at the University of Leon, Spain. Her research interests include learning theory, statistical pattern recognition and neural networks.