Feature Selection Stability Assessment Based on the Jensen-Shannon Divergence*

Roberto Guzmán-Martínez[1] and Rocío Alaiz-Rodríguez[2]

[1] Servicio de Informatica y Comunicaciones
Universidad de León, 24071 León, Spain
roberto.guzman@unileon.es
[2] Dpto. de Ingeniería Eléctrica y de Sistemas,
Universidad de Leon, 24071 León, Spain
rocio.alaiz@unileon.es

Abstract. Feature selection and ranking techniques play an important role in the analysis of high-dimensional data. In particular, their stability becomes crucial when the feature importance is later studied in order to better understand the underlying process. The fact that a small change in the dataset may affect the outcome of the feature selection/ranking algorithm has been long overlooked in the literature. We propose an information-theoretic approach, using the Jensen-Shannon divergence to assess this stability (or robustness). Unlike other measures, this new metric is suitable for different algorithm outcomes: full ranked lists, partial sublists (top-k lists) as well as the least studied partial ranked lists. This generalized metric attempts to measure the disagreement among a whole set of lists with the same size, following a probabilistic approach and being able to give more importance to the differences that appear at the top of the list. We illustrate and compare it with popular metrics like the Spearman rank correlation and the Kuncheva's index on feature selection/ranking outcomes artificially generated and on an spectral fat dataset with different filter-based feature selectors.

Keywords: Feature selection, feature ranking, stability, robustness, Jensen-Shannon divergence.

1 Introduction

Feature selection techniques play an important role in classification problems with high dimensional data [6]. Reducing the data dimensionality is a key step in these applications as the size of the training data set needed to calibrate a model grows exponentially with the number of dimensions (the curse of the dimensionality problem) and the process of knowledge discovery from the data is simplified if the instances are represented with less features.

Feature selection techniques measure the importance of the features according to the value of a given function. These algorithms can be basically divided in

* This work has been partially supported by the Spanish MEC project DPI2009-08424.

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part I, LNAI 6911, pp. 597–612, 2011.
© Springer-Verlag Berlin Heidelberg 2011

three types [7]: 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 model. The wrapper approaches make use of the predictive performance of the classification machine in order 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. Regarding the outcome of the feature selection technique, the output format may be a full ranked list (or weighting-score) or a subset of features. Obviously representation changes are possible and thus, a feature subset can be extracted from a full ranked list by selecting the most important features and a partial ranked list can be also derived directly from the full ranking by removing the least important features.

A problem that arises in many practical problems, in particular when the available dataset is small and the feature dimensionality is high, is that small variations in the data lead to different outcomes of the feature selection algorithm. Perhaps the disparity among different research findings has made the study of the stability (or robustness) of feature selection a topic of recent interest. Fields like biomedicine, bioinformatics or chemometrics require not only accurate classification models, but a feature ranking or 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 (or the ranked feature list) varies, makes this task not straightforward and the conclusions derived from it quite unreliable.

The assessment of the robustness of feature selection/ranking methods becomes an important issue [11,9,3,1], specially when the aim is to gain insight into the underlying process by analyzing the most relevant features. Nevertheless, this is a topic that has received little attention and it has been only during the last decade that several works address this analysis. In order to measure the stability, suitable metrics for each output format of the feature selection algorithms are required.

The Spearman's rank correlation coefficient [10,11,19] and Canberra distance [9] have been proposed to measure the similarity when the outcome representation is a full ranked list. When the goal is to measure the similarity between top-k lists (partial lists), a wide variety of measures have been proposed: Jaccard distance [11,19], an adaptation of the Tanimoto distance [11], Kuncheva's stability index [13], Relative Hamming distance [5], Consistency measures [20], Dice-sorense's index [15], Ochiai's index [22] or Percentage of overlapping features [8]. An alternative that lies between full ranked lists (all features with ranking information) and partial lists (a subset with the top-k features, where all of them are given the same importance) is the use of partial ranked lists, that is, a list with the top-k features and the relative ranking among them. This approach has been used in the information retrieval domain [2] to evaluate queries and it seems more natural when the goal is to analyze a subset of features. Providing information of the feature importance is fundamental to carry

out a subsequent analysis of the data, but a stability measures have not been proposed yet for these partial ranked lists.

In our context, the evaluation of the robustness of feature selection techniques, two ranked lists would be considered much less similar if their differences occurred at the "top" rather than at the "bottom" of the lists. Unlike metrics such as the Kendall's tau and the Spearman's rank correlation coefficient that do not capture this information, we propose a stability measure based on information theory that takes this into consideration. Our proposal is based on mapping each ranked list into a probability distribution and then, measuring the dissimilarity among these distributions using the information-theoretic Jensen-Shannon divergence. This single metric, S_{JS} (Similarity based on the Jensen-Shannon divergence) applies to full ranked lists, partial ranked lists as well as top-k lists.

The rest of this paper is organized as follows: Next, Section 2 describes the stability problem and common approaches to deal with it. The new metric based on the Jensen-Shannon divergence S_{JS} is presented in Section 3. Experimental evaluation is shown in Section 4 and finally Section 5 summarizes the main conclusions.

2 Problem Formulation

In this section, we formulate the problem mathematically and present two common metrics to evaluate the stability of a feature selection and a feature ranking algorithm.

2.1 Feature Selection and Ranking

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 l-dimensional vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots x_{il})$ 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 selection algorithm whose output is a vector $\mathbf{s}$

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

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

When the algorithm performs feature ranking, its output is a ranking vector $\mathbf{r}$ with components

$$\mathbf{r} = (r_1, r_2, r_3, \ldots, r_l) \tag{2}$$

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

Converting a ranking output into a top-k list is conducted according to

$$s_i = \begin{cases} 1 & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

2.2 Similarity Measures

Generally speaking, the dissimilarity among ranked lists can be measured at different levels:

- Among full ranked lists
- Among partial sublists (top-k lists)
- Among partial ranked lists (top-k ranked lists)

Ideally, this metric should be bounded by constants that do not depend on the size of the sublist k or on the total number of features l. Additionally, it should have a constant value for randomly generated subsets/rankings.

The Spearman's rank correlation coefficient is the most frequent metric to measure the similarity between two full ranking outputs [11,19,16,8] but it is not suitable to partial lists. Among the wide variety of metrics proposed to measure the similarity between partial lists with k features, such as the Jaccard distance, Relative Hamming distance, Consistency measures, Dice-sorense's index [15], Ochiai's index [22], Percentage of overlapping features, the Kuncheva's stability index appears to be the most widely accepted [13,1,8]. These metrics only apply to top-k lists, though. Finally, other measures can be applied to full ranked lists and top-k lists: Canberra distance, Spearman's footrule and Spearman's rho, but they do not fulfil desirable properties such as, they should be bounded by constants that do not depend on k or l and it should have a constant value for randomly extracted subsets or random rankings.

Let $\mathbf{r}$ and $\mathbf{r}'$ be the outcome of a feature ranking algorithm applied to two different subsamples of $\mathcal{D}$. The Spearman's rank correlation coefficient (S_R) is defined as

$$S_R(\mathbf{r}, \mathbf{r}') = 1 - 6 \sum_{i=1}^{l} \frac{(r_i - r_i')^2}{l(l^2 - 1)} \tag{4}$$

where l is the number of features. The S_R metric takes values in the interval $[-1, 1]$, being -1 for exactly inverse orders, 0 if there is no correlation between the rankings and 1 when the two rankings are identical.

Let also consider $\mathbf{s}$ and $\mathbf{s}'$ as the outcome of a feature selection algorithm applied to two different subsamples of $\mathcal{D}$. The Kuncheva's index (KI) is given by

$$KI(\mathbf{s}, \mathbf{s}') = \frac{ol - k^2}{k(l - k)} \tag{5}$$

where l is the original whole number of features, o is the number of features that are present in both lists simultaneously and k is the length of the sublists, that is, $\sum_{i=1}^{l} s_i = \sum_{i=1}^{l} s_i' = k$. The KI satisfies $-1 < KI \leq 1$, achieving its maximum when the two lists are identical ($o = k$) and its minimum for independently drawn lists $\mathbf{s}$ and $\mathbf{s}'$.

2.3 The Stability for a Set of Lists

When we have a set of outputs from a feature ranking algorithm, $\mathcal{A} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_N\}$, with size N, 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.

$$S(\mathcal{A}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_M(\mathbf{r}_i, \mathbf{r}_j) \tag{6}$$

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

3 Stability Based on the Jensen-Shannon Divergence

We propose a stability measure based on the Jensen-Shannon Divergence able to measure the diversity either among several full ranked lists, among partial ranked lists or among top-k lists. When the ranking is taken into account, the differences at the top of the list would be considered more important than differences at the bottom part, regardless it is a full or a partial list. When we focus on top-k lists, all the features would be given the same importance.

Our approach to measure the stability of feature selection/ranking techniques is based on mapping the output of the feature selection/ranking algorithm into a probability distribution. Then, the "distance" among these distributions is measured with the Jensen-Shannon divergence [14].

Next, we present our proposal for full ranked lists and then, Section 3.1 and Section 3.2 cover its extension to partial ranked lists and top-k lists, respectively.

Given the output of a feature ranking algorithm, features at the top of the list should be given the highest probability (or weight) and it should smoothly decrease according to the rank. Thus, following [2] the ranking vector $\mathbf{r} = (r_1, r_2, r_3, \ldots, r_l)$ would be mapped into the probability vector $\mathbf{p} = (p_1, p_2, p_3, \ldots, p_l)$ where

$$p_i = \frac{1}{2l}\left(1 + \frac{1}{r_i} + \frac{1}{r_i + 1} + \ldots + \frac{1}{l}\right) \tag{7}$$

where $\sum_{i=1}^{l} p_i = 1$

This way, we assess the dissimilarity between two ranked lists $\mathbf{r}$ and $\mathbf{r}'$, measuring the divergence between the distributions $\mathbf{p}$ and $\mathbf{p}'$ associated with them.

When it comes to measure the difference between two probability distributions, the Kullback-Leibler (KL) divergence D_{KL} [12] becomes the most widely used option. The KL divergence between probability distributions $\mathbf{p}$ and $\mathbf{p}'$ is given by

$$D_{KL}(\mathbf{p}||\mathbf{p}')) = \sum_i p_i \log \frac{p_i}{p'_i} \tag{8}$$

This measure is always non negative, taking values from 0 to ∞, and $D_{KL}(p||q)) = 0$ if $p = q$. The KL divergence, however, has two important drawbacks, since (a) in general it is asymmetric ($D_{KL}(p||q)) \neq D_{KL}(q||p)$) and (b) it does not generalize to more than two distributions. For this reason, we use the related Jensen-Shannon divergence [14], that is a symmetric version of the Kullback-Leibler divergence and is given by

$$D_{JS}(\mathbf{p}||\mathbf{p}') = \frac{1}{2}\left(D_{KL}(\mathbf{p}||\bar{\mathbf{p}}) + D_{KL}(\mathbf{p}'||\bar{\mathbf{p}})\right) \tag{9}$$

where $\bar{\mathbf{p}}$ is the average of the distributions.

Given a set of N distributions $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$, where each one corresponds to a run of a given feature ranking algorithm, we can use the Jensen Shannon divergence to measure the similarity among the distributions produced by different runs of the feature ranking algorithm, what can be expressed as

$$D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(\mathbf{p}_i||\bar{\mathbf{p}}) \tag{10}$$

or alternatively as

$$D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{l} p_{ij} log\frac{p_{ij}}{\bar{p_i}} \tag{11}$$

with p_{ij} being the probability assigned to feature i in the ranking output j and $\bar{p_i}$ the average probability assigned to feature i.

We look for a stability measure based on the Jensen Shannon Divergence (S_{JS}) that fulfills some constraints:

- It falls in the interval [0 ,1]
- It takes the value zero for completely random rankings
- It takes the value one for stable rankings

The stability metric S_{JS} (Stability base on the Jensen Shannon divergence) is given by

$$S_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = 1 - \frac{D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)}{D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N)} \tag{12}$$

where D_{JS} is the Jensen Shannon Divergence among the N ranking outcomes and D^*_{JS} is the divergence value for a ranking generation that is completely random.

In a random setting, $\bar{p_i} = 1/l$ what leads to a constant value D^*_{JS}

$$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{l} p_{ij} log(p_{ij} l) = \frac{1}{N} N \sum_{i=1}^{l} p_i log(p_i l) = \sum_{i=1}^{l} p_i log(p_i l) \tag{13}$$

where p_i is the probability assigned to a feature with rank r_i. Note that this maximum value depends exclusively on the number of features and it can be computed beforehand with the mapping provided by (7).

It is easy to check that:

- For a completely stable ranking algorithm, $p_{ij} = \bar{p_i}$ in (11). That is, the rank of feature-j is the same in any run-i of the feature ranking algorithm. This leads to $D_{JS} = 0$ and a stability metric $S_{JS} = 1$

- A random ranking will lead to $D_{JS} = D^*_{JS}$ and therefore $S_{JS} = 0$
- For any ranking neither completely stable nor completely random, the similarity metric $S_{JS} \in (0, 1)$. The closer to 1, the more stable the algorithm is.

3.1 Extension to Partial Ranked Lists

The similarity between partial ranked lists, that is, partial lists that contain the top-k features with relative ranking information can be also measured with the S_{JS} metric. In this case, the probability is assigned to the top-k ranked features according to

$$p_i = \begin{cases} \frac{1}{2k}\left(1 + \frac{1}{r_i} + \frac{1}{r_i + 1} + \ldots + \frac{1}{k}\right) & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The S_{JS} is computed according to (12) with the normalizing factor D^*_{JS} given by (13) and the probability p_i assigned to a feature with rank r_i computed as stated in (14).

3.2 Extension to Top-k Lists

When it comes to partial lists with a given number k of top-features, a uniform probability is assigned to the selected features according to

$$p_i = \begin{cases} \frac{1}{k} & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

The S_{JS} is computed according to (12) with the probability p_i assigned to a feature according to (15) and the normalizing factor D^*_{JS} given by

$$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_N) = \sum_{i=1}^{l} p_i log(p_i l) = \sum_{i=1}^{k} \frac{1}{k} log\left(\frac{1}{k} l\right) = log\left(\frac{l}{k}\right) \tag{16}$$

where k is the length of the sublist and l the total number of features.

4 Empirical Study

4.1 Illustration on Artificial Outcomes

In this experiment we evaluate the stability metric S_{JS} for the outcomes of hypothetical feature ranking algorithms. We generate sets of $N = 100$ rankings of $l = 2000$ features. We simulate several feature ranking (FR) algorithms:

- FR-0 with 100 random rankings, that is, a completely random FR algorithm
- FR-1 with one fixed output, and 99 random rankings.
- FR-2 with two identical fixed outputs, and 98 random rankings.

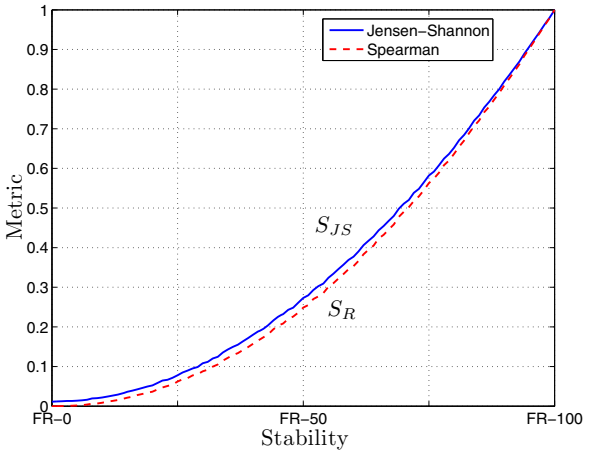

Fig. 1. S_{JS} metric and Spearman rank correlation for Feature Ranking (FR) techniques that vary from completely random (FR-0 on the left) to completely stable (FR-100 on the right)

- FR-i with i identical fixed outputs, and $100 - i$ random rankings.
- FR-100 with 100 identical rankings, that is, an stable FR technique.

Fig. 1 shows the S_{JS} and the S_R for Feature Ranking (FR) techniques that vary from completely random (FR-0, on the left) to completely stable (FR-100 on the right). For the FR-0 method, the stability metric based on Jensen-Shannon divergence S_{JS} takes the value 0, while its value is 1 for the stable FR-100 algorithm. Note that S_{JS} takes similar values to the Spearman rank correlation coefficient S_R.

Assume we have now some Feature Selection (FS) techniques, which stability needs to be assessed. These FS methods (FS-0,FS-1,...,FS-100) have been obtained from the corresponding FR techniques described above, extracting the top-k features ($k = 600$). In the same way, they vary smoothly from a completely random FS algorithm (FS-0) to stable FS a completely stable one (FS-100). The Jensen-Shannon metric S_{JS} together with the Kuncheva Index (KI) are depicted for top-600 lists in Fig. 2. Note that the S_{JS} metric applied to top-k lists provide similar values to the KI metric. The Jensen-Shannon based measure S_{JS} can be applied to full ranked lists and partial lists, while the KI is only suitable to partial lists and the S_R only to full ranked lists.

Generating partial ranked feature lists is an intermediate option between: (a) generating and comparing full ranked feature lists that are, in general, very long and (b) extracting sublists with the top-k features, but with no information about the importance of each feature. The S_{JS} metric based on the Jensen-Shannon divergence also allows to compare these partial ranked lists (as well as top-k lists and full ranked lists). Consider we have sets of sublists with the 600 most important features out of 2000 features. We generated several sets of lists:

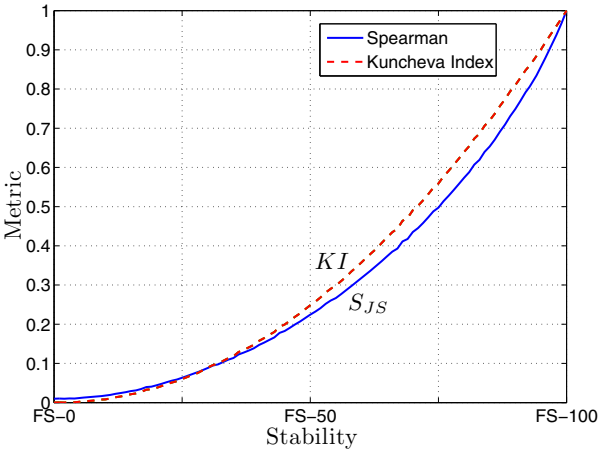

Fig. 2. S_{JS} metric and the KI for Feature Selection (FS) techniques that vary from completely random (FS-0 on the left) to completely stable (FS-100 on the right). The metrics work on top-k lists with k=600.

some of them show high differences in the lowest ranked features while other show high differences in the highest rank features. The same sublist can come either with the ranking information (partial ranked lists) or with no information about the feature importance (top-k lists). The overlap among the lists is around 350 features. Fig. 3 shows the value S_{JS} (partial ranked lists), S_{JS} (top-k list) and the Kuncheva index (top-k lists) for the lists.

Even though the lists have the same average overlap (350 features), some of them show more discrepancy about which are the top features (Fig. 3, on the right), while other sets show more differences at the bottom of the list. The KI can not handle this information since it only works with top-k lists and therefore, it assigns the same value for these very different situations. When the S_{JS} works at this level (top-k list), it also gives the same measure for all the scenarios. The S_{JS} can also handle the information provided in partial ranked lists, considering the importance of the features and therefore assigning a lower stability value for those sets of lists with high differences at the top of the lists, that is with high discrepancy about the most important features. Likewise, it assigns a higher stability value for those sets where the differences appear in the least important features, but there is more agreement about the most important features. Fig. 3 illustrates this fact where S_{JS} (partial ranked lists) varies according to the location of the differences in the list, while S_{JS} (top-k lists) and the KI assign the same value regardless of where the discrepancies appear.

Consider also the following situation where the most important 600 features out of 2000 have been extracted and the overlap among the top-600 lists of 100%. We have evaluated several scenarios:

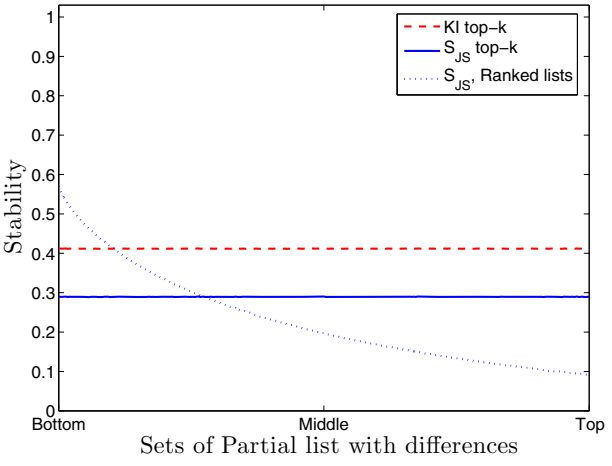

Fig. 3. S_{JS} (partial ranked lists), S_{JS} (top-k list) and the Kuncheva index (top-k lists) for Feature Selection (FS) techniques that extract the top-600 features out of 2000. The overlap among the lists is around 350 common features. The situations vary smoothly from sets of partial lists with differences at the bottom of the list (left) to sets of lists that show high differences at the top of the list (right).

- The feature ranks are identical in all the lists (Identical)
- The ranking of a given feature is assigned randomly (Random)
- Neither completely random nor completely identical.

Working with top-k lists (KI), the stability metrics provide a value of 1, what is somewhat misleading considering the different scenarios that may appear. It seems natural that, even though all agree about the 600 most important features, the stability metric should be lower than 1 when there is low agreement about which are the most important features. The S_{JS} measure allows to work with partial ranked lists and therefore establishing differences between these scenarios. Fig. 4 shows the S_{JS} (partial ranked lists) and the S_{JS}, KI (top-k lists) highlighting this fact. S_{JS} (partial ranked lists) takes a value slightly higher than 0.90 for a situation where there is complete agreement about which are the most important 600 features, but complete discrepancy about their importance. Its value increases to 1 as the randomness in the feature ranking assignment decreases. In contrast with this, KI would assign a value of 1 what may misleading when studying the stability issue.

4.2 Evaluation on an Spectral Dataset

The new measure has been used to experimentally assess the stability of four standard feature selectors based on a filter approach: χ^2, Information Gain Ratio (GR) [4], Relief and other based on the parameter values of an independent classifier (Decision Rule 1R) [21].

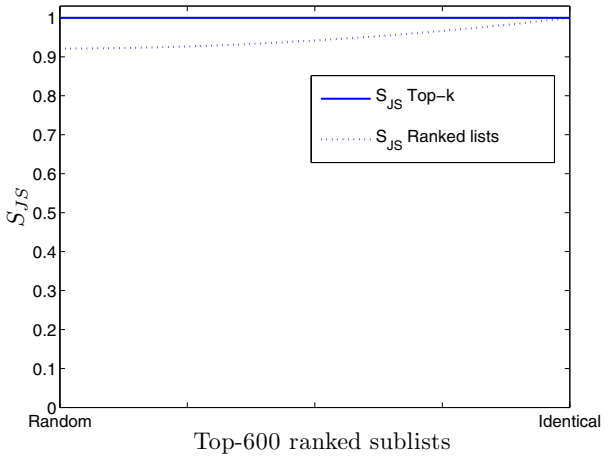

Fig. 4. S_{JS} (top-k list) and S_{JS} (partial ranked lists) for Feature Selection (FS) techniques that extract the top-600 features out of 2000. The overlap among the sublists with 600 features is complete. The ranking assigned to each feature varies from FS techniques for which it is random (left) to FS techniques for which each feature ranking is identical in each sublist (right).

We have conducted some experiments on a real data set of omental fat samples collected from carcasses of suckling lambs [18]. The 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.5.

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 $N = 50$ rankings. Feature ranking was carried out with WEKA [21] and the computation of the stability with MATLAB [17].

The S_{JS} (full ranked list) measure gives an overall view of the stability. The results (Table 1) indicate that in the case of the spectral data, the most stable methods seem to be Relief and GR, while 1R appears as the one with less global stability.

The metric S_{JS} also enables an analysis focused on the top ranked or selected features. Fig. 6 depicts the S_{JS} for a given number of the top-k selected features (continuous line) and the S_{JS} for the the top-k ranked features (dashed line).

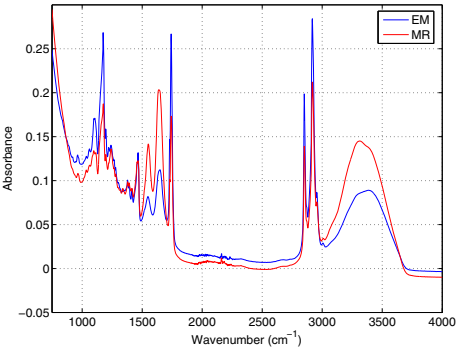

Fig. 5. Average FT-IR spectrum of omental fat samples for Milk Replacer (MR) and Ewe Milk (EM)

Table 1. Stability of several feature selectors evaluated with the similarity measure based on the Jensen-Shannon divergence (S_{JS}) on a set of 50 rankings

S_{JS} (*full ranked list*)			
1R	χ^2	GR	Relief
0.87	0.92	0.94	0.94

The differences between S_{JS} for top-k lists and top-k ranked lists is explained by the fact that in the latter, differences/similarities in the lowest ranks are attached less importance than differences/similarities in the highest ranks. Thus, results show that the four feature selectors share a common trend: S_{JS} (top-k) assigns a lower value of stability that may be sometimes substantially different. Thus, for the 1R feature selector, S_{JS} (ranked top-400) is 0.82, but it drops to 0.70 when all features are given a uniform weight. This is explained by the fact that many differences appear at the bottom of the list and when they are given the same importance as differences at the top of the list, the stability measure drops considerably.

When we focus on the top-k (selected/ranked) features and the value of k is low, the feature selectors are quite stable. For example, for $k = 10$, S_{JS} takes the value 0.92 for χ^2, 0.73 for 1R, 0.92 for GR and 0.91 for Relief.

The plots in Fig. 6 allow to see that the stability decreases as the cardinality of the feature subset increases for the feature selection techniques 1R, χ^2 and GR while Relief shows an stability profile with high stability regardless of the size of sublist. While looking at the whole picture GR is as stable as Relief in general terms, when we focus on lists with the most important features, Relief's robustness does not decrease as the feature subset size increases.

The proposed metric S_{JS} can be compared with the Spearman's rank correlation coefficient (S_R) when it comes to measure the stability of full ranked lists. Likewise, it can be compared with the Kuncheva's stability index (KI) if partial lists are considered. Note, however, that S_{JS} is suitable for whatever problem.

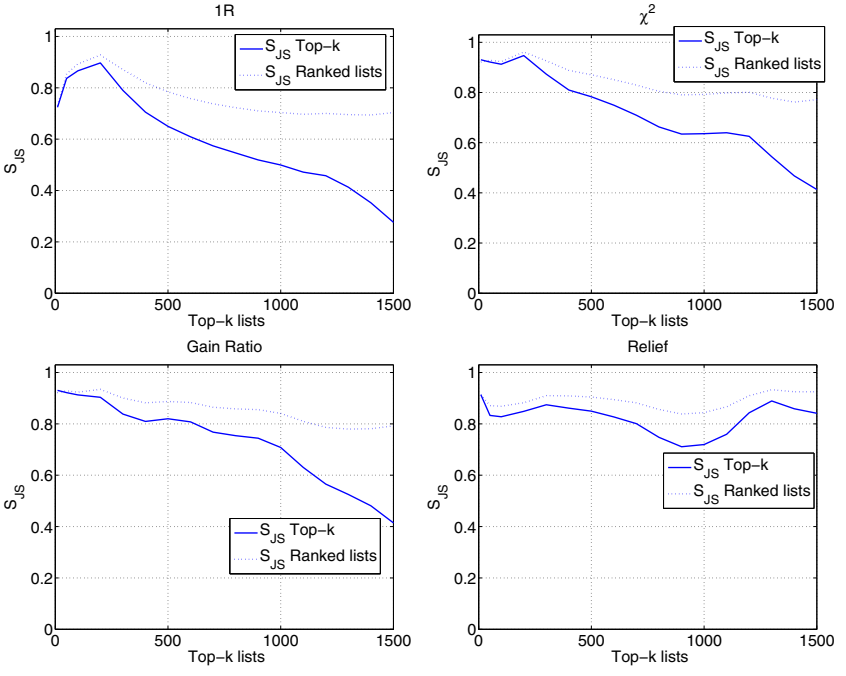

Fig. 6. Feature selection methods 1R, χ^2, GR and Relief applied on the Omental Fat Spectra Dataset. Stability measure S_{JS} for feature subsets with different cardinality.

Table 2. Stability of several feature selectors evaluated with the similarity measure based on the Spearman's rank correlation coefficient (S_R) on a set of 50 rankings.

S_R (*full ranked list*)			
1R	χ^2	GR	Relief
0.79	0.85	0.90	0.94

Measuring the robustness with S_R and KI requires the computation of $\frac{2}{50(50-1)}$ pairwise similarities for each algorithm to end up averaging these computations as stated in Eq.(6). According to the S_R values recorded in Table 2, Relief appears as the most stable (0.94) ranking algorithm, whereas 1R is quite unstable (0.79). When S_{JS} works on the full ranked lists, it gives a stability indication similar to S_R and the findings derived from them are not contradictory. When S_{JS} works on the top-k lists, its value is similar to the provided by KI (see Fig. 7), what allows to see the S_{JS} measure as a generalized S_{JS} metric that can work not only with full ranked lists or top-k lists, but also with top-k ranked lists, while the others are restricted to a particular list format.

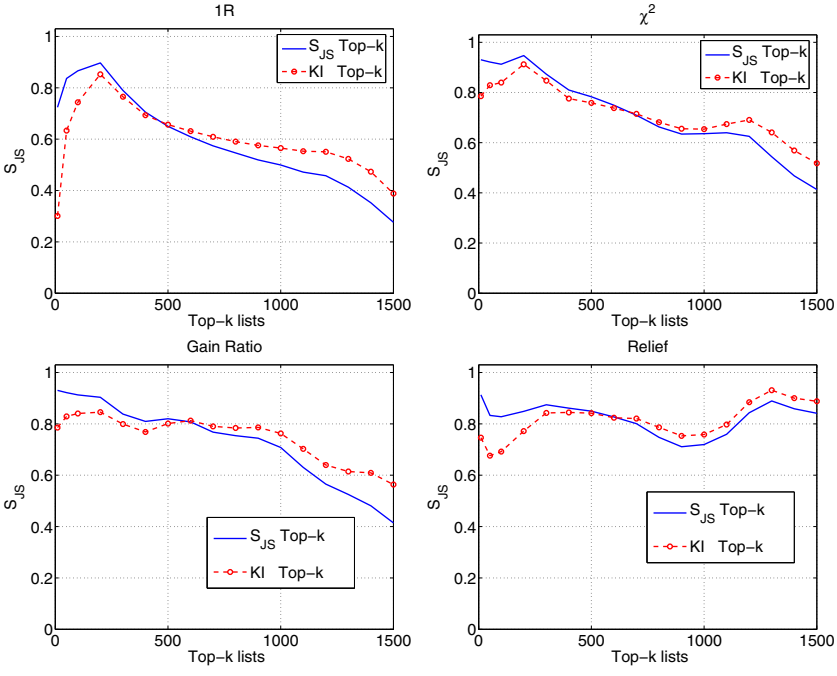

Fig. 7. Feature selection methods 1R, χ^2, GR and Relief applied on the Omental Fat Spectra Dataset. Stability measure S_{JS} and KI for top-k lists with different cardinality.

5 Conclusions

The robustness of the feature ranking techniques used for knowledge discovery is an issue of recent interest. In this work, we consider the problem of feature selection/ranking stability and propose a metric based on the Jensen-Shannon divergence (S_{JS}) able to capture the disagreement among the lists generated in different runs by a feature ranking algorithm from different perspectives: (a) considering the full ranked feature lists, (b) focusing on the top-k features, that is to say, lists that contain the k most relevant features giving a uniform importance to all them and (c) considering partial ranked lists that retain the most relevant features together with the ranking information.

The new metric S_{JS} shows the relative amount of randomness of the ranking/selection algorithm, independently of the sublist size and unlike other metrics that evaluate pairwise similarities, S_{JS} evaluates directly the whole set of lists (with the same size). Up to our knowledge, no metrics have been proposed so far to measure the similarity between partial ranked feature lists. Moreover, the new measure accepts whatever representation of the feature selection output and its behavior is: (i) close to the Spearman's rank correlation coefficient for full

ranked lists and (ii) similar to the Kuncheva's index for top-k lists. If the ranking is taken into account, the differences at the top of the list would be considered more important than differences that appear at the bottom part.

The stability of feature selection algorithms opens a wide area of research that includes the development of more robust feature selection techniques, their correlation with classifier performance and different approaches to analyze robustness. The proposal of visual techniques to ease the stability analysis and the exploration of committee-based feature selectors is our immediate future research.

References

1. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3), 392 (2010)
2. Aslam, J., Pavlu, V.: Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 198–209. Springer, Heidelberg (2007)
3. Boulesteix, A.-L., Slawski, M.: Stability and aggregation of ranked gene lists 10(5), 556–568 (2009)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley and Sons, Chichester (2001)
5. Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Trinity College Dublin Computer Science Technical Report, 2002–2028
6. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)
7. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature Extraction: Foundations and Applications. Studies in Fuzziness and Soft Computing. Springer-Verlag New York, Inc., Secaucus (2006)
8. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Technical Report arXiv:1001.0887 (January 2010)
9. Jurman, G., Merler, S., Barla, A., Paoli, S., Galea, A., Furlanello, C.: Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics 24(2), 258 (2008)
10. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms. In: Fifth IEEE International Conference on Data Mining, p. 8. IEEE, Los Alamitos (2005)
11. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems 12, 95–116 (2007), doi:10.1007/s10115-006-0040-8
12. Kullback, S., Leibler, R.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951)
13. Kuncheva, L.I.: A stability index for feature selection. In: Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, pp. 390–395. ACTA Press (2007)
14. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991)

15. Loscalzo, S., Yu, L., Ding, C.: Consensus group stable feature selection. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 567–576 (2009)
16. Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S.: Measuring Stability of Feature Selection in Biomedical Datasets. In: AMIA Annual Symposium Proceedings, vol. 2009, p. 406. American Medical Informatics Association (2009)
17. MATLAB. version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts (2010)
18. Osorio, M.T., Zumalacrregui, J.M., Alaiz-Rodrguez, R., Guzman-Martnez, R., Engelsen, S.B., Mateo, J.: 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 83(1), 140–147 (2009)
19. Saeys, Y., Abeel, T., Peer, Y.: Robust Feature Selection Using Ensemble Feature Selection Techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008)
20. Somol, P., Novovicova, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1921–1939 (2010)
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)
22. Zucknick, M., Richardson, S., Stronach, E.A.: Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. Statistical Applications in Genetics and Molecular Biology 7(1), 7 (2008)