Knowledge-Based Systems 195 (2020) 105745

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

An information theoretic approach to quantify the stability of feature selection and ranking algorithms☆

Rocío Alaiz-Rodríguez [a,*], Andrew C. Parnell [b]

[a] Department of Electrical, Systems and Automatic Engineering, Universidad of León. Campus de Vegazana s/n, 24071 León, Spain
[b] Hamilton Institute, Maynooth University, Maynooth, Ireland

ARTICLE INFO

Article history:
Received 10 December 2019
Received in revised form 18 February 2020
Accepted 6 March 2020
Available online 12 March 2020

Keywords:
Feature selection
Feature ranking
Stability
Robustness
Jensen–Shannon divergence

ABSTRACT

Feature selection is a key step when dealing with high-dimensional data. In particular, these techniques simplify the process of knowledge discovery from the data by selecting the most relevant features out of the noisy, redundant and irrelevant features. A problem that arises in many of these practical applications is that the outcome of the feature selection algorithm is not stable. Thus, small variations in the data may yield very different feature rankings. Assessing the stability of these methods becomes an important issue in the previously mentioned situations. We propose an information-theoretic approach based on the Jensen–Shannon divergence to quantify this robustness. Unlike other stability measures, this metric is suitable for different algorithm outcomes: full ranked lists, feature subsets as well as the lesser studied partial ranked lists. This generalized metric quantifies the difference among a whole set of lists with the same size, following a probabilistic approach and being able to give more importance to the disagreements that appear at the top of the list. Moreover, it possesses desirable properties including correction for change, upper/lower bounds and conditions for a deterministic selection. We illustrate the use of this stability metric with data generated in a fully controlled way and compare it with popular metrics including the Spearman's rank correlation and the Kuncheva's index on feature ranking and selection outcomes, respectively. Additionally, experimental validation of the proposed approach is carried out on a real-world problem of food quality assessment showing its potential to quantify stability from different perspectives.

1. Introduction

Feature selection is a key step in many classification problems [1–3], in particular in those with high dimensional datasets. It is well known that the size of the training dataset needed to calibrate a model grows exponentially with the number of dimensions (the curse of the dimensionality problem). The main motivation to implement these techniques has been to improve the classification performance by selecting an optimum subset of features. Numerous papers have examined feature selection with respect to classification performance [4–6].

Additionally, the process of knowledge discovery from the data in fields like biomedicine, bioinformatics, genetics or chemometrics is simplified with the use of feature selection methods. Removing the noisy and irrelevant features while keeping the most relevant features is essential for understanding the underlying process.

Identifying the most relevant features for the problem studied has been the goal of many research papers. It has been applied to discriminate different types of cancer [7,8], to categorize healthy and diseased tissue [9], to uncover the risk factors for a disease [10,11] or to select genes related to a disease [12–14].

Although feature selection techniques are of great help to identify the most relevant features in these domains, a problem that arises in many practical problems is that the outcome of the feature selection algorithm does not tend to be stable in the sense that small variations in the data may yield to very different feature rankings.

Stability (or robustness) issues have long been overlooked in the literature. However, the topic of robustness of feature selection techniques has attracted an increasing interest in the machine learning field in the past few years [15–24]. The issues have arisen as a consequence of the difficulties of reproducing different research findings.

Evaluating the stability of the lists that come out of the feature ranking (or selection) techniques becomes crucial before trying to gain insight into the data. Otherwise, the conclusions derived

☆ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper..

* Corresponding author.
E-mail addresses: rocio.alaiz@unileon.es (R. Alaiz-Rodríguez), Andrew.Parnell@mu.ie (A.C. Parnell).

https://doi.org/10.1016/j.knosys.2020.105745
0950-7051/

from the study may be completely unreliable. Suitable stability metrics for the different outcomes of the feature selection algorithms are required. Moreover, these metrics should possess desirable properties so that they allow for a useful interpretation of stability and similarly comparisons among feature selection methods.

Most proposals to quantify stability only apply to feature subsets: Jaccard distance [15,25], Tanimoto distance [15], Kuncheva's stability index [26], Consistency measures [27], etc. Others only deal with full ranked lists such as the Spearman's rank correlation coefficient [15,25,28] or the Canberra distance [16].

An interesting alternative that lies between full ranked lists and lists with feature subsets 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 [29] to evaluate queries and it seems more natural when the goal is to analyze a subset of features.

It seems reasonable that when it comes to assess the robustness of feature selection techniques, two ranked lists should be considered much less similar if their differences occurred at the "top" rather than at the "bottom" of the lists. Up to our knowledge only a modified version of the Canberra distance has been proposed for this purpose [30] but it does not extend to other feature selection outcomes.

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} (Stability based on the Jensen–Shannon divergence) applies to different algorithm outcomes: full ranked lists, partial ranked lists as well as top-k lists with equal length. Furthermore, it also fulfills the desirable properties for a stability metric so that it enables suitable comparison and interpretation of stability values.

The rest of this paper is organized as follows: In Section 2 we formulate the problem of feature selection. In Section 3 we describe the robustness issue and common metrics to quantify the stability together with a comparison among them. The new metric based on the Jensen–Shannon divergence S_{JS} is presented in Section 4. Experimental evaluation is shown in Section 5. Finally Section 6 summarizes our main conclusions.

2. Feature selection techniques

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

Feature selection techniques measure the importance of a feature ranking or a subset of features according to a given measure [1,31]. From a functional point of view the output of a feature selection algorithm may be a ranking (weighting-score) on the features or feature set [5,31]. 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 relevant features.

Consider now a feature ranking algorithm that leads to a ranking vector $\mathbf{r}$ with components

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

where $1 \leq r_i \leq t$. Note that 1 is considered the highest rank.

Consider also a top-k list as the outcome of a feature selection technique

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

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

Lists with a full ranking of features can be converted into top-k lists that contain the most important k features. Converting a ranking output into a feature subset is easily conducted according to

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

A fundamental property of a feature selection method is its robustness [19,24,32]. This becomes critical in many domains where the stability of a feature selection method is crucial for interpretation by domain experts. Robustness has been defined as the sensitivity of the method to small perturbations in the training set [15].

Suppose we ran a feature ranking algorithm K times and obtained a set of rankings $\mathcal{A} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_K\}$. For the purpose of illustration, Fig. 1 shows an example where instances are defined by ten features ($t = 10$) and the feature ranking algorithm is applied to five different subsamples of the data ($K = 5$)

Once obtained, the dissimilarity among the outputs can be measured at different levels:

- Among full ranked lists
- Among feature subsets (top-k lists)
- Among partial ranked lists (top-k ranked lists)

Thus, the outcomes for full ranked lists can be gathered in a matrix $\mathcal{A}$ with elements r_{ij} with $i = 1, \ldots, t$ and $j = 1, \ldots, K$ that indicate the rank assigned in run j for feature i. Note that $\mathcal{A}_{fr}$ stands for set of lists with full ranking.

$$\mathcal{A}_{fr} = \begin{bmatrix} \mathbf{r}'_1 & \mathbf{r}'_2 & \mathbf{r}'_3 & \mathbf{r}'_4 & \mathbf{r}'_5 \end{bmatrix} = \begin{bmatrix} 3 & 9 & 7 & 8 & 7 \\ 2 & 1 & 2 & 3 & 3 \\ 4 & 7 & 3 & 5 & 2 \\ 9 & 6 & 10 & 9 & 8 \\ 5 & 3 & 5 & 1 & 4 \\ 10 & 5 & 8 & 7 & 9 \\ 7 & 10 & 9 & 10 & 10 \\ 8 & 2 & 6 & 6 & 5 \\ 1 & 4 & 1 & 2 & 1 \\ 6 & 8 & 4 & 4 & 6 \end{bmatrix}_{10\times 5}$$

Fig. 2 shows the top-4 ranked lists and the top-4 lists for the example presented above. Additionally, some of the stability metrics that are commonly applied for each output format are also shown.

The outcomes for the top-4 lists can also be gathered in a matrix $\mathcal{A}_f$ with elements s_{ij} with $i = 1, \ldots, t$ and $j = 1, \ldots, K$ that indicate whether or not the feature-i has been selected among the top-4 most relevant in the run-j.

$$\mathcal{A}_s = \begin{bmatrix} \mathbf{s}'_1 & \mathbf{s}'_2 & \mathbf{s}'_3 & \mathbf{s}'_4 & \mathbf{s}'_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}_{10\times 5}$$

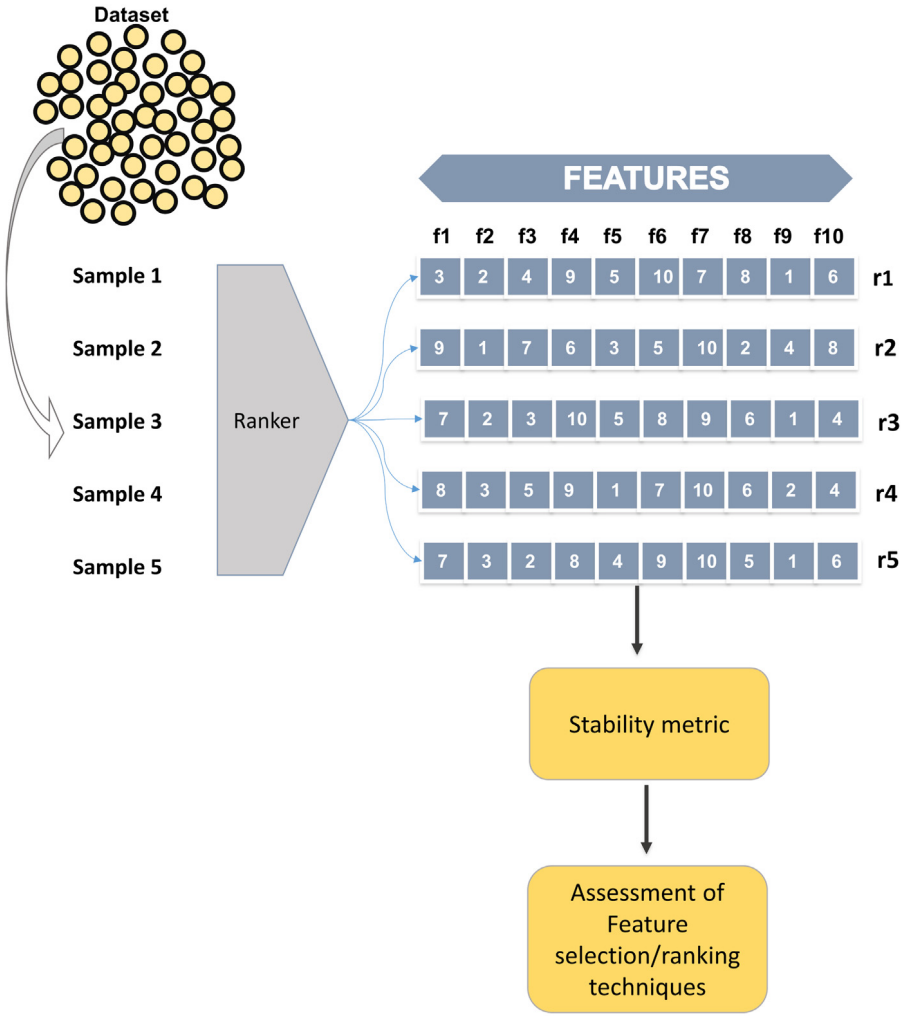

Fig. 1. Illustration of the stability problem for feature ranking methods.

In the case of dealing with partial ranked lists, the set of lists can be represented in matrix $\mathcal{A}_{pr}$

$$\mathcal{A}_{pr} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 2 & 1 & 2 & 3 & 3 \\ 4 & 0 & 3 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 1 & 4 & 1 & 2 & 1 \\ 0 & 0 & 4 & 4 & 0 \end{bmatrix}_{10\times 5}$$

In next section, we review the most common ways to assess the robustness of feature selection techniques.

3. Related work: Quantifying the stability of feature selection techniques

Non-stability of feature selection is a problem that may appear in practical applications, but in particular it is more noticeable when the available dataset is small and the feature dimensionality is high, as is common in biomedicine, bioinformatics, and chemometrics. Instability issues make the feature rankings unreliable for clinical use. Therefore, it becomes essential to provide metrics to evaluate the robustness of given feature selection techniques when applied to our data. Efforts have also been made in order to increase the robustness of feature selection methods [21,22,33–35].

The assessment of stability has attracted great interest in the field of feature selection. In general, stability is quantified following two different approaches: (i) Given a set of rankings (or subsets), pairwise similarities are computed and then reduced to a single metric by averaging. (ii) Defining a function applied on matrix $\mathcal{A}$ but not based on pairwise similarities. Another qualitative option for this assessment is a visual analysis of stability.

Next, we present desirable properties for the stability metrics and a review of previous proposals highlighting their advantages and drawbacks.

3.1. Properties of stability metrics

There are some properties that a stability metric should possess so that it allows for a useful interpretation of stability and similarly comparisons among feature selection/ranking techniques. Kuncheva [26] was the first to provide a list of desirable properties for a similarity measure S_M between two feature subsets of equal length.

These properties proposed in [26], however, only refer to similarity measures (S_M) and they do not necessarily imply that the stability index obtained by computing pairwise similarities have the same properties as the distance metric. On the other hand, there are other proposals such as [27] (or ours) that are not

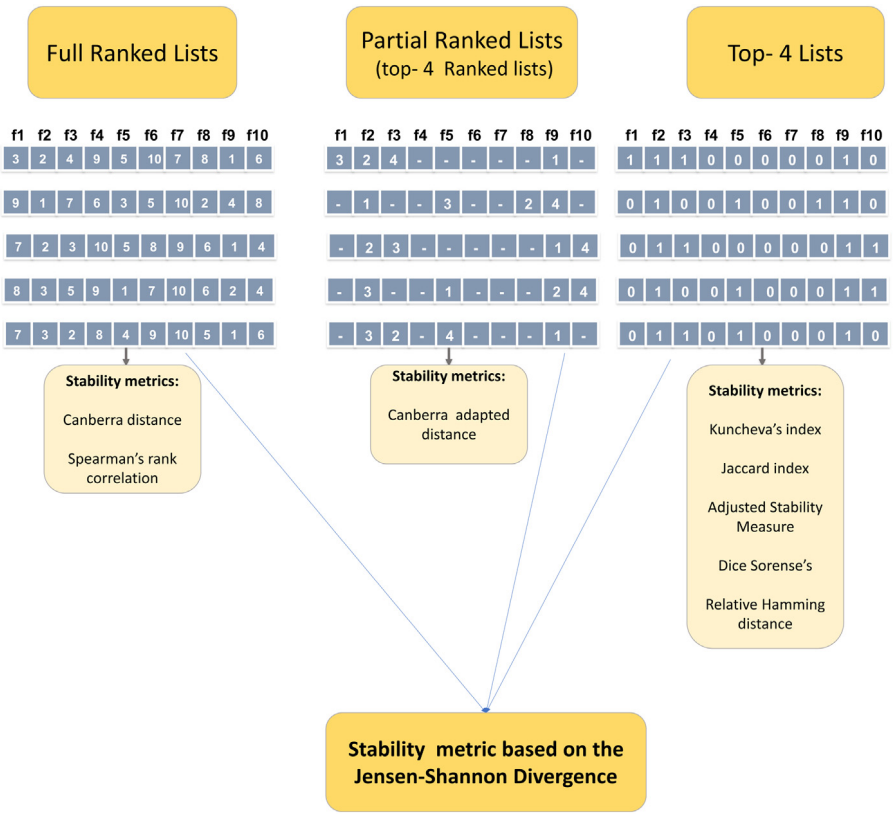

Fig. 2. Outcomes for feature selection techniques: full ranked lists, partial ranked lists (top-k ranked lists) and top-k lists.

based on calculating pairwise similarities. These properties were later refined in [24] where the authors study the properties from the wider viewpoint of the stability metric.

Nogueira et al. [24] focused on feature selection techniques that may select feature subsets of arbitrary cardinality identifying some properties necessary for a given stability measure. These desirable properties are: upper and lower bounds, correction for chance, maximum stability, and fully defined. They further showed that many stability measures widely used in the literature do not possess all these properties.

The four properties proposed by [26] and [24] are:

Property 1: *Upper and Lower Bounds*
The stability metric Φ should have upper and lower bounds that do not depend on the total number of features or the feature subset length.
Property 2: *Maximum* $\longleftrightarrow$ *Deterministic Selection* The stability metric $\Phi(\mathcal{A})$ should reach its maximum if-and-only-if all feature sets in $\mathcal{A}$ are identical.
Property 3: *Correction For Chance*
When the selection is random, that is feature sets of size k_i have an equal probability of being drawn, the expected value of $\Phi(\mathcal{A})$ should be constant, which is set for convenience to 0.
Property 4: *Fully Defined*
The stability metric $\Phi(\mathcal{A})$ should be completely defined for any set $\mathcal{A}$ of features. This property ensures the stability metric can cope with feature subsets of any size.

This latter property enables the application of a stability metric to domains where the feature selection algorithm may return subsets with different number of features. Nonetheless, we consider this property as optional but not essential since there are many scenarios in which the number of selected features is fixed to a number k for a given study.

3.2. Stability metrics based on computing pairwise similarities

The most widely used approach to evaluate the stability of a feature selection (or ranking) algorithm that provides several results $\mathcal{A} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_K\}$, is to compute pairwise similarities and average the results. This approach leads to a scalar value:

$$\Phi(\mathcal{A}) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} S_M(\mathbf{r}_i, \mathbf{r}_j) \tag{3}$$

where S_M refers to any similarity metric which takes as input the appropriate format of $\mathcal{A}$.

3.2.1. Similarity metric for feature subsets

When the goal is to measure the similarity between feature subsets (also referred as top-k lists) different authors have proposed similarity metrics: Jaccard distance [15], Tanimoto distance [15], Kuncheva's stability index [26], Adjusted Stability Measure (ASM) [36], Relative Hamming distance [37], Dice–Sorense's index [38], Ochiai's index or Percentage of overlapping features [39]. Of these, the Jaccard distance and the Kuncheva's stability index appear to be the most widely accepted [22,26,39].

Let consider now $\mathbf{s}$ and $\mathbf{s}'$ as the output vector of a feature selection algorithm applied to two different subsamples of $\mathcal{D}$.

The Jaccard stability index is defined as

$$JI(\mathbf{s}, \mathbf{s}') == \frac{|\mathbf{s} \wedge \mathbf{s}'|}{|\mathbf{s} \vee \mathbf{s}'|} = \frac{o}{l} \tag{4}$$

where $\mathbf{s}$ and $\mathbf{s}'$ are the two feature subsets, o is the number of features that are common in both lists and l the number of features that appear in any of the two lists. The Jaccard index lies in the range (0, 1). This metric can cope with feature subsets of different length but it does not take into account the similarity between

Table 1
Eligible stability metrics for different feature rankings and feature subset formats.

Stability metric	Full ranked lists	Partial ranked lists	Partial ranked lists with different length	Feature subset lists	Feature subset lists with different length
Canberra distance [40]	Yes	–	–	–	–
Canberra adapted distance [30]	Yes	Yes	Yes	–	–
Spearman's rank correlation coefficient [15]	Yes	–	–	–	–
Jaccard distance [15]	–	–	–	Yes	Yes
Tanimoto [15]	–	–	–	Yes	Yes
Relative Hamming distance [37]	–	–	–	Yes	Yes
Kuncheva's stability index [26]	–	–	–	Yes	–
Adjusted Stability Measure ASM [36]	–	–	–	Yes	Yes
Relative Weighted Consistency CWrel [27]	–	–	–	Yes	Yes
Dice–Sorense's index [38]	–	–	–	Yes	Yes
Our proposal: Jensen–Shannon stability metric	Yes	Yes	–	Yes	–

subsets of features due to chance (randomness) and therefore it does not possess the correction for chance property [24].

The Kuncheva's index (KI) for these two top-k lists is given by

$$KI(\mathbf{s}, \mathbf{s}') = \frac{ot - k^2}{k(t-k)} \tag{5}$$

where t is the total number of features, o is the number of features that are present in both lists and k is the length of the sublists, that is, $\sum_{i=1}^{t} s_i = \sum_{i=1}^{t} s'_i = k$. The KI satisfies $-1 < KI \leq 1$, achieving its maximum when the two lists are identical ($o = k$) and values close to zero for independently drawn lists $\mathbf{s}$ and $\mathbf{s}'$ (i.e. o expected to be around k^2/t).

The KI metric possesses the aforementioned three properties of a stability metric but it is limited to feature subsets of equal length and it cannot be extended to ranked lists (either full or partial).

3.2.2. Similarity metric for ranked lists (full and top-k)

Consider $\mathbf{r}$ and $\mathbf{r}'$ the output of a feature ranking technique applied to two subsamples of $\mathcal{D}$. The Spearman's rank correlation coefficient [15,25,36,39] and Canberra distance [40] have been proposed to measure the similarity between rankings. Of the two, Spearman's rank correlation coefficient (S_R) is perhaps the most popular. The S_R between two ranked lists $\mathbf{r}$ and $\mathbf{r}'$ is defined by

$$S_R(\mathbf{r}, \mathbf{r}') = 1 - 6 \sum_{i=1}^{t} \frac{(r_i - r'_i)^2}{t(t^2-1)} \tag{6}$$

where r_i is the rank of feature-i and t the total number of features. S_R values range from -1 to 1. It takes the value one when the rankings are identical and the value zero when there is no agreement between rankings. This metric is only suitable for lists with the same size.

A less studied situation is to focus on the top ranked features but unlike the top-k list, keeping the ranking information. Thus, the Canberra distance, initially proposed to assess the similarity between full feature rankings, was extended to top-k ranked lists using a location parameter [30].

3.3. Stability metrics based on a function definition

Generally, we can define a function $\Phi(\mathcal{A})$ to avoid computing all pairwise similarities. A popular measure in this category is the Relative Weighted Consistency Measure CWrel [27]. This stability metric is a direct function of the frequency of the features after feature selection. Other proposals within this category include the frequency of selection normalized by the number of feature subsets and averaged over all features [23]. Although the CWrel metric can cope with lists with arbitrary length, it lacks of two desired aforementioned properties: maximum and correction for change (proofs can be found in [24]).

3.4. Metric comparison: Advantages and limitations

The output of a technique that selects the most relevant features may come in different formats. Table 1 summarizes for which output format some widely known stability metrics can be applied.

It is evident that the stability metrics developed for feature subsets cannot deal with rankings and in general the opposite is also true. There are metrics, however, such us the Canberra distance – initially proposed for full feature rankings – that was extended to compute the distance between upper partial lists of the original rankings [30]. By contrast, we propose a stability measure, S_{JS}, that can deal with full ranked lists, partial ranked lists as well as feature sets.

Regarding the three properties mentioned previously, Table 2 shows a general overview of these properties for several stability metrics proposed for feature selection methods. Note that the fully defined property defined in [24] was included in Table 2 (last column, feature subsets with arbitrary length). As mentioned before, we consider it useful to determine whether or not a stability metric can be applied to a given output format but in our opinion this cannot be viewed as an essential property by itself.

Many of the metrics proposed for computing the stability for feature sets can handle lists of different length: Tanimoto, ASM, CWrel, the relative Hamming distance, the Jaccard distance and Dice–Sorense's index. The popular Kuncheva index (and ours) applies only to lists with equal length, though. Note, however,

Table 2
Properties of stability metrics for feature selection methods.

Stability metric	Upper and Lower bounds	Maximum	Correction
Jaccard distance [15]	Yes	Yes	–
Tanimoto [15]	Yes	Yes	–
Relative Hamming distance [37]	Yes	Yes	–
Kuncheva's stability index [26]	Yes	Yes	Yes
Adjusted Stability Measure ASM [36]	Yes	–	Yes
Relative Weighted Consistency CWrel [27]	Yes	–	–
Dice–Sorense's index [38]	Yes	Yes	–
Our proposal: Jensen–Shannon stability	Yes	Yes	Yes

that these metrics able to deal with lists with arbitrary length do not possess either one or two of the desired properties.

Thus, some well known stability metrics do not verify the *correction for change* property, such as Tanimoto, Relative Hamming distance, Jaccard distance, Dice–Sorense's index or the CWrel metric. Besides this, stability metrics like ASM or CWrel do not fulfill the *Maximum* property. It is only the Kuncheva index that has the three properties that appear in Table 2.

The main strength of the metric S_{JS} we propose is that it possesses as we will show in Section 4.3 the essential properties defined for a stability metric and additionally, it can deal either with feature rankings (full /partial) or feature subsets. Moreover, it does not require such pairwise comparison that becomes computationally expensive for large datasets. In any case, the feature lists should have the same number of elements.

3.5. Visual analysis of robustness

The outcome of a feature ranking algorithm can be interpreted as a point in a high dimensional space (with t dimensions). The stability of a pairwise ranking can be viewed as computing distances between points in that high dimensional space and averaging the results. These (scalar) metrics can be seen as projections to one dimensional space and their use only provides guidance as to where the feature selector stands in relation to a stable and a random ranking algorithm.

The use of graphical methods as a simple alternative approach to evaluate the stability of feature ranking algorithms has been proposed in [11]. It has been highlighted that if we change from a projection to a space with one dimension, into a space with two or more dimensions, we can conduct a visual analysis that allows the user to visually assess stability as well as establish comparisons with other feature ranking or selection methods.

In [11], a dimensionality reduction technique like Multi-Dimensional Scaling (MDS) [41] has been proposed for a visual analysis of robustness. It allows the projection of data from a high dimensional space to a 2D or 3D space while preserving the distance in the original high dimensional space.

Fig. 3 illustrates this approach with several feature ranking algorithms:FR-a, FR-b, FR-c, FR-d, FR-e. The algorithms are run on seven sub-samples of the data. This figure allows the user to see in a single figure that the most unstable algorithm is FR-a since the points are very scattered. The outcomes of FR-d, however, are clustered together. The same applies to the FR-e and these therefore are the most stable. This figure also allows the user to see that FR-e generates a similar ranking to FR-d. Finally, note that FR-c is very different to the aforementioned groups.

4. An information theoretic approach to measure robustness

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 between these distributions is measured with the Jensen–Shannon divergence [42].

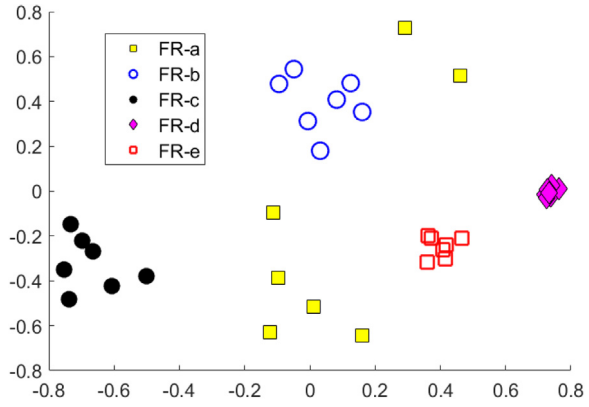

Fig. 3. Visual-based stability analysis for five hypothetical feature rankings (FR).

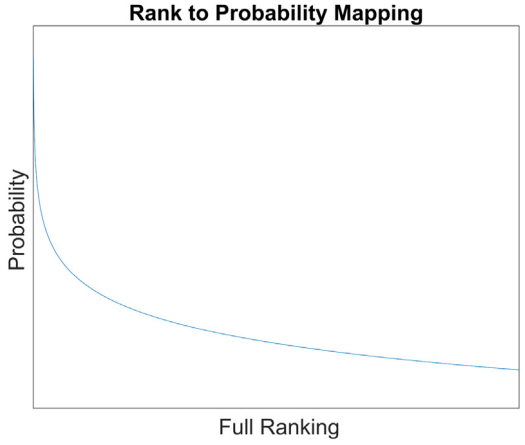

Fig. 4. Mapping of ranks into probabilities for full ranked lists.

Below we present our proposal for full ranked lists and then Sections 4.1 and 4.2 describe its extension to top-k ranked lists and lists with feature subsets, 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 [29] the ranking vector $\mathbf{r} = (r_1, r_2, r_3, \ldots, r_t)$ would be mapped into the probability vector $\mathbf{p} = (p_1, p_2, p_3, \ldots, p_t)$ where

$$p_i = \frac{1}{2t}\left(1 + \sum_{j=0}^{t-r_i}(r_i + j)^{-1}\right) \tag{7}$$

where by design $\sum_{i=1}^{t} p_i = 1$. Fig. 4 illustrates the mapping of rankings into probabilities for full ranked lists.

We can thus quantify the similarity between two ranked lists $\mathbf{r}$ and $\mathbf{r}'$ by measuring the divergence between the distributions $\mathbf{p}$ and $\mathbf{p}'$ associated with them.

The most widely used metric for measuring the difference between two probability distributions is the Kullback–Leibler (KL) divergence D_{KL} [43], 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 ($D_{KL}(p\|q) \neq D_{KL}(q\|p)$ thus not a true distance measure, and (b) it does not generalize to more than two distributions. For this reason, we use the related Jensen–Shannon divergence [42], 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 K distributions $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K\}$, 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

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

or equivalently as

$$D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K) = \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{t} p_{ij} \log \frac{p_{ij}}{\bar{p}_i} \tag{11}$$

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.

Some desirable constraints that this stability measure possesses includes:

- It falls in the interval [0 ,1]
- It takes the value zero for completely random rankings
- It takes the value one for stable rankings
- It is invariant to the ordering of the ranking probability distributions

We define the stability metric S_{JS} (Stability based on the Jensen–Shannon divergence) as:

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

where D_{JS} is the Jensen–Shannon Divergence among the K ranking outcomes and D^*_{JS} is the divergence value for a ranking generation that is completely random. In a random setting, $\bar{p}_i = 1/t$ which leads to a constant value D^*_{JS}

$$\begin{aligned} D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K) &= \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{t} p_{ij} \log(p_{ij}t) \\ &= \frac{1}{K}K\sum_{i=1}^{t} p_i \log(p_i t) = \sum_{i=1}^{t} p_i \log(p_i t) \end{aligned} \tag{13}$$

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).

We can 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.

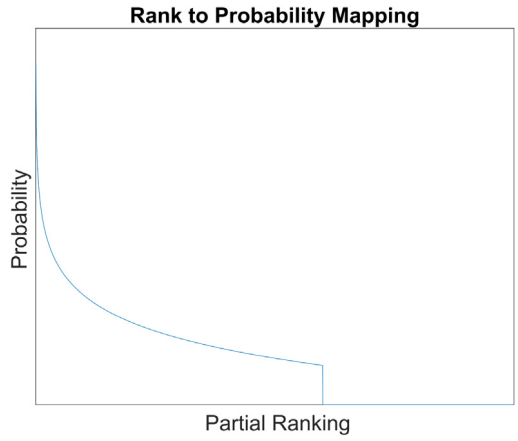

Fig. 5. Mapping of ranks into probabilities for partial ranked lists.

4.1. 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 is:

$$p_i = \begin{cases} \dfrac{1}{2k}\left(1 + \displaystyle\sum_{j=0}^{k-r_i}(r_i + j)^{-1}\right) & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Fig. 5 shows the mapping of rankings into probabilities for top-k ranked lists.

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).

4.2. Extension to feature subsets

When we deal with feature subsets with a given number of top-k features, a uniform probability is assigned to the selected features (see Fig. 6) according to

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

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

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

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

Algorithm 1 summarizes the different options for computing the stability metric S_{JS} depending on the outcome of the feature selection technique.

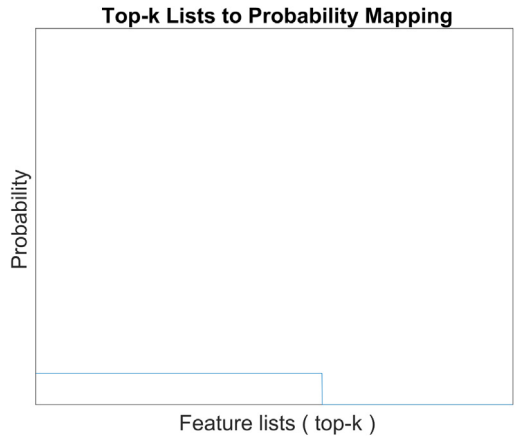

Fig. 6. Mapping of ranks to probabilities for feature subsets (top-k lists).

Algorithm 1: Stability metric based on the Jensen–Shannon Divergence S_{JS}

h! function S_{JS} ($\mathcal{A}$, *ListFormat*);
Input : Matrix $\mathcal{A}$: Feature selection/ranking outputs
ListFormat: List Format in $\mathcal{A}$
Output: $S_{JS}(\mathcal{A})$
if *ListFormat=FullRanking* **then**
$p_i = \frac{1}{2t}\left(1 + \sum_{j=0}^{t-r_i}(r_i + j)^{-1}\right)$
$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K) = \sum_{i=1}^{t} p_i \log(p_i t)$
else if *ListFormat=PartialRanking* **then**
$p_i = \begin{cases} \frac{1}{2k}\left(1 + \sum_{j=0}^{k-r_i}(r_i + j)^{-1}\right) & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases}$
$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K) = \sum_{i=1}^{t} p_i \log(p_i t)$
else if *ListFormat=FeatureSubset* **then**
$p_i = \begin{cases} \frac{1}{k} & \text{if } r_i \leq k \\ 0 & \text{otherwise} \end{cases}$
$D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K) = \log\left(\frac{t}{k}\right)$
end
Compute $D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K) = \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{t} p_{ij} \log \frac{p_{ij}}{\bar{p}_i}$
Return $S_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K) = 1 - \frac{D_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K)}{D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K)}$

4.3. Properties of the stability metric S_{JS}

The S_{JS} stability measure presented in this work possesses the first three of the aforementioned properties:

Property 1: *Upper and Lower Bounds*
The stability metric S_{JS} takes values in the interval [0 ,1]
Property 2: *Maximum* $\longleftrightarrow$ *Deterministic Selection*
The stability metric S_{JS} reaches its maximum value 1 if-and-only-if all feature sets in $\mathcal{A}$ are identical.
Property 3: *Correction For Chance*
When the selection is random, there is a normalizing term $D^*_{JS}(\mathbf{p}_1, \ldots, \mathbf{p}_K)$ that corresponds to the divergence value for a feature set or ranking that is completely random. In that case, S_{JS} takes the value 0.

The stability metric we propose in this paper focuses on problems where the feature rankings or feature subsets have the same length, hence the *fully defined* property does not apply.

To summarize, feature selection techniques can represent feature relevance as a feature subset or as a ranking. So far, there exist stability metrics that are only suitable for specific outcomes of the feature selection techniques. Our proposal provides a solution for unifying stability of subset selection and rankings since S_{JS} applies to feature subsets and full ranked lists. It also extends to the least studied partial ranked lists that only keep the ranking information for the best ranked features. This is interesting since the variability of the ranks for non relevant features definitely adds noisy information in this process.

When the ranking is taken into account (either full or partial ranked lists) differences at the top of the list would be considered more important than differences at the bottom part. This unifying framework satisfies the desired properties for a stability metric (bounds, maximum and correction for change). Moreover, it is function-based metric that, unlike most stability metrics, is not based on computing pairwise similarities, and therefore is less computationally expensive for large high-dimensional datasets. The application of the S_{JS} metric is limited, however, to lists with the same cardinality.

The S_{JS} metric has been initially proposed to measure the stability of feature selection techniques. Nonetheless, it can be extended to other applications that require ranking comparisons since this metric provides an indication of how similar rankings are. For instance, ranking is a very important topic in information search where retrieval rankings need to be compared in order to find, for example, patterns of similarity and differences in sets of rankings. Other domains like computational biology, genetics or biomedicine also base decision-making processes on sets of rankings that need to be compared.

5. Experimental results

5.1. Illustration on artificial outcomes

In this section we illustrate the stability metric S_{JS} for the outcomes of some hypothetical feature ranking algorithms. We generate sets of $N = 100$ rankings of $t = 2000$ 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.
- FR-i with i identical fixed outputs, and $100 - i$ random rankings.
- FR-100 with 100 identical rankings, that is, a stable FR technique.

Fig. 7 shows our stability metric based on the Jensen–Shannon divergence (S_{JS}) compared ton the Spearman's rank correlation coefficient (S_R) for 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 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's rank correlation coefficient S_R.

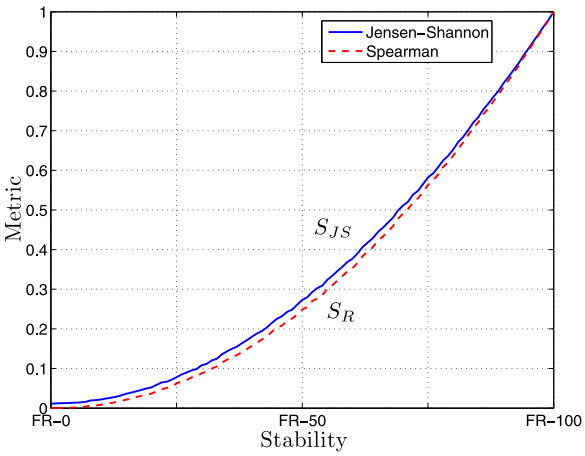

Fig. 7. S_{JS} metric and Spearman rank correlation for Feature Ranking (FR) techniques that vary from completely random (FR-0 on the left) to completely stable (FR-100 on the right).

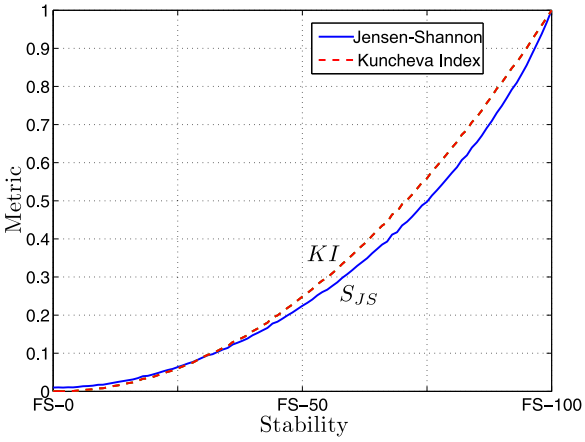

Fig. 8. S_{JS} metric and the KI for Feature Selection (FS) techniques that vary from completely random (FS-0 on the left) to completely stable (FS-100 on the right). The metrics work on top-k lists with k=600.

Suppose now we have some Feature Selection (FS) techniques, for 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. 8. Note that the S_{JS} metric applied to top-k lists provides 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 for partial lists and the S_R only to full ranked lists.

Generating partial ranked feature lists is an intermediate step 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 relevance information for each feature. The S_{JS} metric based on the Jensen–Shannon divergence also allows to compare these partial ranked lists.

Suppose we have sets of sublists with the 600 most important features out of 2000 features. We generated several sets of lists: some of them show high differences in the lowest ranked features whilst others show high differences in the highest rank features. The same sublist can come either with the ranking information

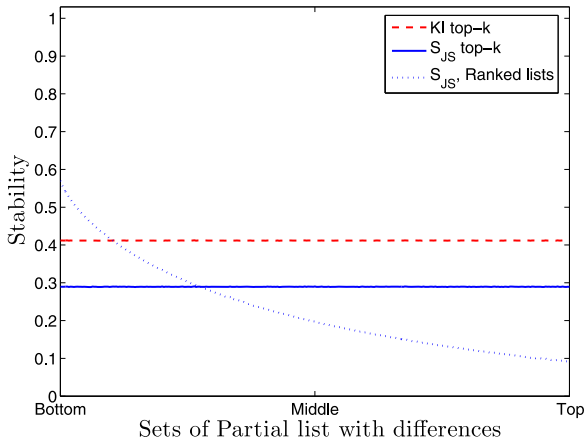

Fig. 9. S_{JS} (partial ranked lists), S_{JS} (top-k list) and the Kuncheva index (top-k lists) for Feature Selection (FS) techniques that extract the top-600 features out of 2000. The overlap among the lists is around 350 common features. The situations vary smoothly from sets of partial lists with differences at the bottom of the list (left) to sets of lists that show high differences at the top of the list (right).

(partial ranked lists) or with no information about the feature importance (top-k lists). The overlap among the lists is around 350 features. Fig. 9 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. 9, on the right), while other sets show more differences at the bottom of the list.

The KI cannot 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. 9 illustrates this fact where S_{JS} (for 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.

Next, consider the situation where the most important 600 features out of 2000 have been extracted and the overlap among the top-600 lists is 100%. We have evaluated several scenarios:

- 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 that 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 us to work with partially ranked lists and therefore establish differences between these scenarios. Fig. 10 shows the S_{JS} (partial ranked lists) and the S_{JS}, KI (top-k lists) highlights this fact. S_{JS} (partial ranked lists) takes a value

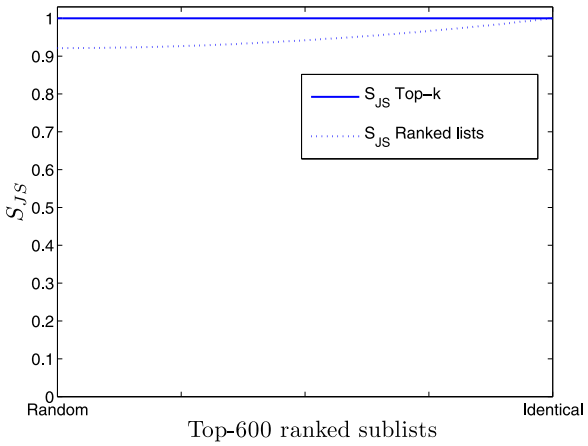

Fig. 10. S_{JS} (top-k list) and S_{JS} (partial ranked lists) for Feature Selection (FS) techniques that extract the top-600 features out of 2000. The overlap among the sublists with 600 features is complete. The ranking assigned to each feature varies from FS techniques for which it is random (left) to FS techniques for which each feature ranking is identical in each sublist (right).

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 which may mislead when studying the stability issue.

5.1.1. Computational time

In order to more thoroughly examine the stability metrics and consider real-time constraints, we have measured the execution time required to compute the stability metrics analyzed previously in four different scenarios. These scenarios, that differ in the number of features t and the number of feature lists K, are:

- Scenario-A: $K = 100$ and $t = 2000$
- Scenario-B: $K = 300$ and $t = 2000$
- Scenario-C: $K = 100$ and $t = 3000$
- Scenario-D: $K = 300$ and $t = 3000$

The procedures have been implemented in Matlab and for program profiling the following configuration was used:

- Processor: Intel (R) Core (TM) 2 Duo E4500 (2.20 GHz).
- Memory: 2 GB.
- Operating System: Windows10 Enterprise (64-bit).

Fig. 11 shows execution times for the S_R and S_{JS} metrics in the aforementioned scenarios. As we can see, the computation time for S_R is always higher than the required for S_{JS}, being from twice its value to 24 times higher. Execution time for S_R is also very sensitive to both dimensionality and the number of rankings. Consider we change from scenario-A to scenario-D, where this latter scenario has 200 more rankings and the lists have 1000 more features than scenario-A. In this situation, computation time for KI goes from 60 ms in scenario-A to 1830 ms in scenario-D. The execution time for S_{JS} is, however, much lower and less sensitive as it goes from 27 ms in scenario-A to 75 ms in scenario-D.

Fig. 12 shows execution times for the *KI* and S_{JS} for top-600 lists extracted from the four scenarios defined above. The KI index also requires more computational time than S_{JS} in all the scenarios evaluated. For instance, S_{JS} is computed in 6 ms when we have 100 rankings and 2000 lists of top-600 features while KI requires 39 ms. This means the computation of the KI metric 6 times much slower that the one for S_{JS}. Differences are

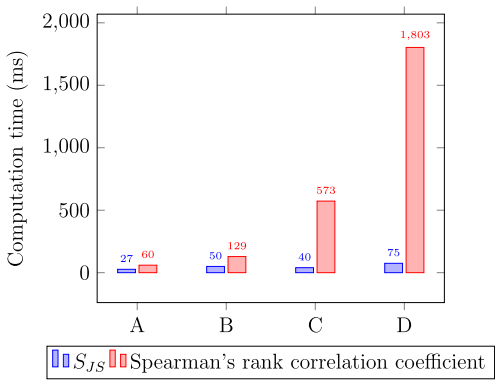

Fig. 11. Computation time (ms) for the S_{JS} stability metric and the Spearman's rank correlation coefficient for different scenarios: Scenario A ($K = 100$, $t = 2000$), scenario B ($K = 300$, $t = 2000$), scenario C ($K = 100$, $t = 3000$), scenario D ($K = 300$, $t = 3000$).

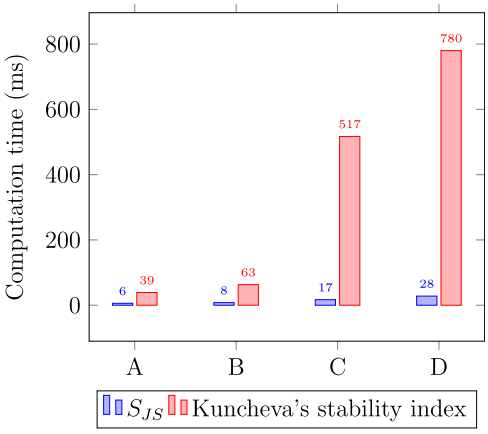

Fig. 12. Computation time (ms) for the S_{JS} stability metric and the Kuncheva's stability index coefficient for different scenarios with top-600 features: Scenario A ($K = 100$, $t = 2000$), scenario B ($K = 300$, $t = 2000$), scenario C ($K = 100$, $t = 3000$), scenario D ($K = 300$, $t = 3000$).

even larger when we increase either the number of lists or the dimensionality. Thus, in Scenario-D, computing S_{JS} takes 28 ms whereas computing KI requires 780 ms, (over 27 times more time than calculating S_{JS}) what makes S_{JS} more suitable for real-time applications.

5.2. Practical application of food quality assessment

In this section we evaluate the S_{JS} metric to quantify the stability of feature selection techniques applied to a real practical application of food quality assessment. We address the problem of authentication of suckling lamb meat with respect to the type of feeding. The rearing system determines the difference in quality and prices in the market. The use of FTIR spectroscopy for the discrimination of fat samples according to the rearing system provides several advantages over conventional analytical methods in a laboratory, mainly its speed, cost and versatility.

The FTIR spectra comprise, however, a large number of irrelevant and redundant information. Appropriate feature selection is an aid to identify spectrum regions with more prediction power and link this information with chemical interpretation, what has high interest for the veterinarian professionals.

Omental fat samples were collected from carcasses of suckling lambs [44]. Lambs came from the flocks of three farms affiliated to the 'Asociación Nacional de Criadores de Ganado Ovino de Raza

Table 3
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

Table 4
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

Churra', which is a Churra breeders association from the region of 'Castilla-Leon' (Spain). Lambs were reared either exclusively on Ewe Milk (EM) or on a Milk Replacer (MR). The whole dataset has 134 instances: 66 from lambs being fed with a MR, while the other 68 are reared on EM. 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 S_{JS} stability metric has been used to experimentally assess the stability of four feature selectors: χ^2 [45–47], Information Gain Ratio (GR) [46,48,49], Relief [46,50] and other based on the parameter values of an independent classifier (Decision Rule 1R) [46,51].

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 [46] and the computation of the stability with MATLAB [52].

The S_{JS} (full ranked list) measure gives an overall view of the stability. The results (Table 3) 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. 13 depicts the S_{JS} for a given number of the top-k selected features (continuous line) and the S_{JS} for the top-k ranked features (dashed line).

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. 13 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. Looking at the whole picture GR is as stable as Relief. However, when we focus on lists with the most important features, GR's robustness decreases as the feature subset size increases, whereas Relief does not.

Next, we compare the proposed metric S_{JS} with the Spearman's rank correlation coefficient (S_R) when it comes to measure the stability of full ranked lists. Likewise, we compare it with the Kuncheva's stability index (KI) if partial lists are considered. Note, however, that S_{JS} is more versatile and suitable for whatever output format.

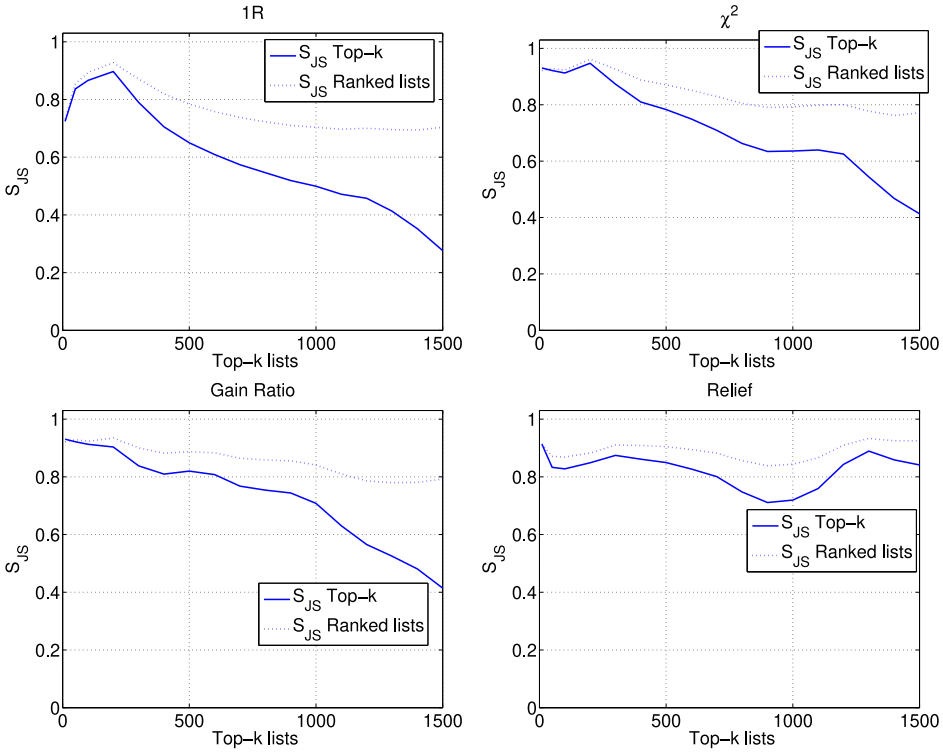

Fig. 13. Feature selection methods 1R, χ^2, GR and Relief applied on the Omental Fat Spectra Dataset. Stability measure S_{JS} for feature subsets with different cardinality.

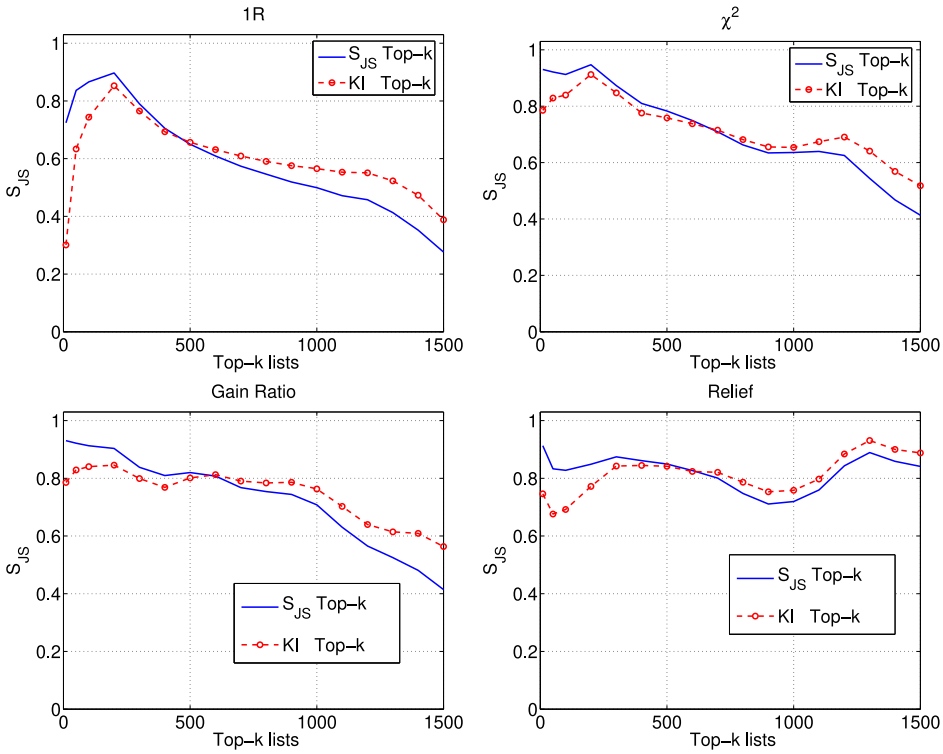

Fig. 14. Feature selection methods 1R, χ^2, GR and Relief applied on the Omental Fat Spectra Dataset. Stability measure S_{JS} and KI for top-k lists with different cardinality.

Measuring the robustness with S_R and *KI* requires the computation of $\frac{50(50-1)}{2}$ pairwise similarities for each algorithm to end up averaging these computations as stated in Eq. (3). According to the S_R values recorded in Table 4, 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 quantification of stability 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. 14), 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.

This study shows that either for feature selection or feature ranking and regardless of the cardinality of the feature list, Relief is the most stable feature selector for this problem.

6. Conclusions

Quantifying the stability of feature selection/ranking techniques becomes crucial before trying to gain insight into the data.

We have proposed a unifying stability metric based on the Jensen–Shannon divergence (S_{JS}) able to quantify the stability for whatever outcome of the feature selection techniques (feature subsets, full rankings as well as the useful, but least studied, partial rankings). Up to our knowledge, no metric has been proposed so far to quantify the stability at all these levels.

The S_{JS} stability metric has the following desired properties for a stability metric: upper and lower bounds, conditions for a deterministic selection and correction for change. Therefore, it enables a useful interpretation of stability as well as comparisons among feature selection/ranking techniques.

Unlike most metrics that are based on computing pairwise similarities, S_{JS} evaluates the whole set of lists directly, what makes this approach faster and more efficient for large high-dimensional datasets. This metric applies to feature sets/rankings with the same length, though.

The experimental study with an artificial dataset generated in a fully controlled way show that the new metric S_{JS} is: (a) close to the Spearman's rank correlation coefficient for full ranked lists, (b) similar to the Kuncheva's index for top-k lists and (c) able to capture the mismatch among sublists with the top-k ranked features. In that sense, it can be seen as a generalized metric.

Experimental results with a real problem of food quality assessment shows that S_{JS} is able to quantify the stability from different perspectives. It allows to see that Relief has a profile with high stability regardless of the number of top relevant features and 1R turns out to be the least stable feature selection method for this practical application.

Potential future work includes the exploration of visual techniques with this new metric embedded and its extension to feature lists with arbitrary length.

CRediT authorship contribution statement

Rocío Alaiz-Rodríguez: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft. **Andrew C. Parnell:** Methodology, Writing - review & editing.

Acknowledgments

This research has been funded with support from the European Commission under the 4NSEEK project with Grant Agreement 821966. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

Andrew Parnell's work was supported by a Science Foundation Ireland Career Development Award grant 17/CDA/4695 and an SFI centre, Ireland grant 12/RC/2289_P2.

References

[1] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, Lotfi A. Zadeh, Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing), Springer-Verlag New York, Inc., Secaucus, NJ, USA, ISBN: 3540354875, 2006.

[2] G. Victo, V. Cyril Raj, Review on feature selection techniques and the impact of SVM for Cancer classification using gene expression profile, CoRR (2011).

[3] M. Al-Rajab, J. Lu, Q. Xu, Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis, Comput. Methods Programs Biomed. 146 (2017) 11–24.

[4] Yvan Saeys, Iñaki Inza, Pedro Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[5] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.

[6] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (6) (2017) 94.

[7] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, Z. Yakhini, Tissue classification with gene expression profiles, in: Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, ACM, 2000, pp. 54–64.

[8] T. R Golub, D. K Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P Mesirov, H. Coller, M.L. Loh, J.s R Downing, M.A. Caligiuri, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537.

[9] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Amer. Statist. Assoc. 97 (457) (2002) 77–87.

[10] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.

[11] N. Cueto-López, M.T. García-Ordás, V. Dávila-Batista, V. Moreno, N. Aragonés, R. Alaiz-Rodríguez, A comparative study on feature selection for a risk prediction model for colorectal cancer, Comput. Methods Programs Biomed. 177 (2019) 219–229.

[12] S. Sayed, M. Nassef, A. Badr, I. Farag, A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets, Expert Syst. Appl. 121 (2019) 233–243.

[13] J.C. Ang, A. Mirzal, H. Haron, H.N.A. Hamed, Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinform. 13 (5) (2015) 971–989.

[14] S. Ahmed, M. Kabir, Z. Ali, M. Arif, F. Ali, D.-J. Yu, An integrated feature selection algorithm for cancer classification using gene expression data, Comb. Chem. High Throughput Screen. 21 (9) (2018) 631–645.

[15] Alexandros Kalousis, Julien Prados, Melanie Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowl. Inf. Syst. (ISSN: 0219-1377) 12 (2007) 95–116, http://dx.doi.org/10.1007/s10115-006-0040-8.

[16] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, C. Furlanello, Algebraic stability indicators for ranked lists in molecular profiling, Bioinformatics 24 (2) (2008) 258–264.

[17] A. Boulesteix, M. Slawski, Stability and aggregation of ranked gene lists, Brief. Bioinform. 10 (5) (2009) 556–568.

[18] H. Wang, T. M. Khoshgoftaar, A. Napolitano, Stability of filter- and wrapper-based software metric selection techniques, in: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, 2014, pp. 309–314.

[19] R. Guzmán-Martínez, R. Alaiz-Rodríguez, Feature selection stability assessment based on the Jensen-Shannon divergence, in: Proceedings of the 2011 ECML-KDD Conference - Volume Part I, Springer-Verlag, 2011, pp. 597–612.

[20] F. Yang, K.Z. Mao, Robust feature selection for microarray data based on multicriterion fusion, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (4) (2011) 1080–1092.

[21] B. Pes, N. Dessı, M. Angioni, Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data, Inf. Fusion 35 (2017) 132–147.

[22] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics (ISSN: 1367-4803) 26 (3) (2010) 392.

[23] W.W.B. Goh, L. Wong, Evaluating feature-selection stability in next-generation proteomics, J. Bioinform. Comput. Biol. 14 (05) (2016) 1650029.

[24] Sarah Nogueira, Gavin Brown, Measuring the stability of feature selection, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2016, pp. 442–457.

[25] Y. Saeys, T. Abeel, Y. Peer, ECML PKDD '08: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II.

[26] L.I. Kuncheva, 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, ACTA Press, 2007, pp. 390–395.

[27] P. Somol, J. Novovicova, Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality, IEEE Trans. Pattern Anal. Mach. Intell. (ISSN: 0162-8828) 32 (11) (2010) 1921–1939, http://dx.doi.org/10.1109/TPAMI.2010.34.

[28] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms, in: Data Mining, Fifth IEEE International Conference on, IEEE, ISBN: 0769522785, 2005, p. 8.

[29] J. Aslam, V. Pavlu, Query hardness estimation using jensen-Shannon divergence among multiple scoring functions, in: Giambattista Amati, Claudio Carpineto, Giovanni Romano (Eds.), Advances in Information Retrieval, in: Lecture Notes in Computer Science, vol. 4425, Springer Berlin / Heidelberg, 2007, pp. 198–209.

[30] G. Jurman, S. Riccadonna, R. Visintainer, C. Furlanello, Algebraic comparison of partial lists in bioinformatics, PLoS One 7 (5) (2012) e36540.

[31] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowl. Inf. Syst. 34 (3) (2013) 483–519.

[32] P. Somol, J. Novovicova, Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1921–1939.

[33] H. Liu, L. Liu, H. Zhang, Ensemble gene selection for cancer classification, Pattern Recognit. 43 (8) (2010) 2763–2772.

[34] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: homogeneous and heterogeneous approaches, Knowl.-Based Syst. 118 (2017) 124–139.

[35] V. Bolón-Canedo, A. Alonso-Betanzos, Ensembles for feature selection: a review and future trends, Inf. Fusion 52 (2019) 1–12.

[36] J.L. Lustgarten, V. Gopalakrishnan, S. Visweswaran, Measuring stability of feature selection in biomedical datasets, in: AMIA Annual Symposium Proceedings, vol. 2009, American Medical Informatics Association, 2009, p. 406.

[37] K. Dunne, P. Cunningham, F. Azuaje, Solutions to Instability Problems with Sequential Wrapper-Based Approaches to Feature Selection, Trinity College Dublin Computer Science Technical Report, Citeseer, 2002.

[38] S. Loscalzo, L. Yu, C. Ding, Consensus group stable feature selection, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, 2009.

[39] Z. He, W. Yu, Stable Feature Selection for Biomarker Discovery, Technical Report, 2010, arXiv:1001.0887.

[40] G. Jurman, S. Riccadonna, R. Visintainer, C. Furlanello, Canberra distance on ranked lists, in: Proceedings of Advances in Ranking NIPS 09 Workshop, 2009, pp. 22–27.

[41] T. Cox, M. Cox, Multidimensional Scaling, Chapman and Hall, 1994.

[42] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inform. Theory (ISSN: 00189448) 37 (1) (1991) 145–151.

[43] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.

[44] M.T. Osorio, J.M. Zumalacárregui, R. Alaiz-Rodrí guez, R. Guzman-Martínez, S.B. Engelsen, 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 Sci. (ISSN: 0309-1740) 83 (1) (2009) 140–147.

[45] H. Liu, R. Setiono, Chi2: Feature selection and discretization of numeric attributes, in: TAI '95: Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, 1995, p. 88, ISBN: 0-8186-7312-5.

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

[47] C.S.unil Kumar, R.J. Sree, Application of ranking based attribute selection filters to perform automated evaluation of descriptive answers through sequential minimal optimization models, ICTACT J. Soft Comput. 5 (1) (2014).

[48] M. A. Hall, L.A. Smith, Practical feature subset selection for machine learning, in: Proc. 21st Australian Computer Science Conference, 1998, pp. 181–191.

[49] R. Praveena Priyadarsini, M.L. Valarmathi, S. Sivakumari, Gain ratio based feature selection method for privacy preservation, ICTACT J. Soft Comput. 1 (4) (2011) 201–205.

[50] Kenji Kira, Larry A. Rendell, A practical approach to feature selection, in: Machine Learning Proceedings 1992, Elsevier, 1992, pp. 249–256.

[51] G. Holmes, C.G. Nevill-Manning, Feature selection via the discovery of simple classification rules, in: Proc Symposium on Intelligent Data Analysis (IDA-95), Baden-Baden, Germany, pp. 75–79.

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