Measuring the Stability of Feature Selection with Applications to Ensemble Methods

Sarah Nogueira$^{(\boxtimes)}$ and Gavin Brown

School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{sarah.nogueira,gavin.brown}@manchester.ac.uk

Abstract. Ensemble methods are often used to decide on a good selection of features for later processing by a classifier. Examples of this are in the determination of Random Forest *variable importance* proposed by Breiman, and in the concept of *feature selection ensembles*, where the outputs of multiple feature selectors are combined to yield more robust results. All of these methods rely critically on the concept of feature selection *stability* - similar but distinct to the concept of *diversity* in classifier ensembles. We conduct a systematic study of the literature, identifying desirable/undesirable properties, and identify a weakness in existing measures. A simple correction is proposed, and empirical studies are conducted to illustrate its utility.

Keywords: Stability · Feature selection · Ensembles

1 Introduction

The stability of feature selection can be seen as its sensitivity to *small changes* in the input dataset. In many applications, stability of feature selection is crucial as the user might need to identify an interpretable feature subset, e.g. when identifying genes responsible for a disease [1]. Stable feature selection frameworks provide more reliable feature subsets and gain in interpretability. As the output of a feature selection algorithm (FSA) can either be a set of features, a ranking on the features or a score on the features, there exist stability measures that apply to each one of these cases. In this paper, we focus on the first case where an FSA returns a feature set.

Why is measuring stability of feature selection an issue? In regression or classification predictors, the sensitivity to changes in data is quantified exactly in a *bias-variance* decomposition of the error measure (though in the classification case this is not entirely straightforward). There is no such decomposition that applies to feature selection. First of all, the *true* relevant set of features is unknown (and strongly depends on the classifier that will be used afterwards) which does not allow us to define the concept of bias. In the case of regression predictors, such decomposition relies on the convexity of the squared-loss function. A stability measure will allow us to quantify the variability in the feature sets selected by an FSA for a given dataset.

Why do we need a new measure? Kuncheva [8] demonstrated the importance of the property of *correction for chance* and derived a new measure satisfying this property. Nevertheless, the measure proposed can only be calculated for FSAs selecting a fixed amount of features on a given dataset. Even though several variants of Kuncheva's measure satisfying the property of correction for chance have been proposed to deal with feature sets of varying cardinality [9,12,14], we will show that they are flawed in the sense that they do not satisfy other critical properties (e.g. they do not always return their maximal value when the FSA always returns the same feature set or they are not bounded by constants). Hence, we derived a generalization of Kuncheva's stability measure that can be used with feature sets of varying cardinality while retaining a set of desirable properties. Examples that illustrate of the utility of these measures are feature selection techniques using hypothesis testing, random forests or LASSO. Indeed, when applying LASSO to different samples of the same data, there is no guarantee that the same coefficients will be equal to 0 and hence, that a constant number of features will be selected when different samples of the data are taken.

Applications to Ensemble-Based Feature Selection. Stability of ensemble-based feature selection has recently become area of interest [1,3,5,10]. In ensemble-based feature selection, we use a set of *diverse* feature selection methods to build a more *robust* one. A stability analysis could then be carried out to observe the diversity (corresponding to low stability) of the different feature selection methods within an ensemble as well as to observe the robustness (corresponding to high stability) of the feature selection made by the ensemble.

   The remainder of the paper is structured as follows. Section 2 presents some of the properties of the existing measures. Section 3 focuses on the measures having the property of correction for chance for feature sets of different cardinalities and highlights their weaknesses on toy examples. Section 4 proposes a new measure having a set of identified properties and Sect. 5 illustrates its utility in the context of an ensemble-based feature selection procedure.

2   Stability Measures

2.1   Existing Measures

To observe the robustness of an FSA to changes in the data, the FSA is applied to K samples of the same dataset to obtain a sequence $\mathcal{A}$ of K feature sets. The more similar these K feature sets will be, the more the procedure will be said to be stable. To define stability, one common approach consists of defining a *similarity measure sim* between two feature sets s_1 and s_2 and then to define the stability as the average similarity over all pairs of feature sets in $\mathcal{A}$. In that case, the stability will be denoted by $\overline{sim}$ and we can express it as follows:

$$\overline{sim}(\mathcal{A}) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} sim(s_i, s_j), \qquad (1)$$

where s_i is the ith feature set in $\mathcal{A}$. Several similarity measures have been proposed in the literature. Some of the older works propose using the *Jaccard index* sim_J [6] (also referred as the *Tanimoto distance*) or the relative *Hamming distance* to define a similarity measure sim_H [4]. Let us assume that we have n features in total. The output of an FSA can then be seen as a binary string of length n with a 1 at the ith position if the ith feature has been selected and with a 0 otherwise. Let's assume that an FSA returns the following sequence $\mathcal{A}$ of $K = 3$ feature sets:

$$s_1 :\ 1\,0\,0\,1\,0\,1 \qquad s_2 :\ 1\,1\,0\,0\,0\,1 \qquad s_3 :\ 1\,0\,1\,1\,1\,1 \qquad (2)$$

The similarity measure sim_H between two feature sets is defined as the number of bits they have in common divided by the length n of the string. Therefore, using Eq. 1, the resulting stability of these feature sets will be equal to:

$$\overline{sim_H}(\mathcal{A}) = \frac{2}{3(3-1)} \sum_{i=1}^{2} \sum_{j=i+1}^{3} sim_H(s_i, s_j) = \frac{2}{3(3-1)} \left(\frac{4}{6} + \frac{4}{6} + \frac{2}{6} \right) = \frac{5}{9}.$$
$$(3)$$

Nevertheless, both these measures are *subset-size-biased* [8] meaning that their values are biased by the number of features selected and hence cannot be used consistently to compare the stability of FSAs in different settings. Indeed, imagine that a procedure selects two identical feature sets of 8 features out of a total of 10 features and that another procedure selects two identical feature sets of 8 features out of a total of 100 features. Intuitively, the second procedure is more stable, as it is less likely to have selected the exact same 8 features by chance. For this reason, Kuncheva [8] proposed a similarity measure having the property of correction for chance. The similarity between two feature sets of size k can be seen as the number of features r they have in common (i.e. the size of their intersection). As we want this measure to reflect on the true ability of the procedure to select identical features, Kuncheva [8] proposes correcting r by the expected size of the intersection between two feature sets of k features drawn at random (denoted hereafter by $\mathbb{E}[r]$). The size of the intersection between two sets containing k objects each individually randomly drawn without replacement amongst a total of n objects follows a hypergeometric distribution, and therefore we have that $\mathbb{E}[r] = \frac{k^2}{n}$. In order to make this value comparable for different values of k and n, Kuncheva rescales $r - \mathbb{E}[r]$ in $[-1, 1]$ by dividing it by its maximal value $max(r - \mathbb{E}[r])$:

$$sim_K(s_1, s_2) = \frac{r - \mathbb{E}[r]}{\max(r - \mathbb{E}[r])} = \frac{r - \mathbb{E}[r]}{\max(r) - \mathbb{E}[r]} = \frac{r - \mathbb{E}[r]}{k - \mathbb{E}[r]} = \frac{r - \frac{k^2}{n}}{k - \frac{k^2}{n}}, \quad (4)$$

where s_1 and s_2 are two feature sets of cardinality k and where $max(r)$ is the maximal possible value of r for a given k. The measure sim_K will hence reach its maximal value of 1 when $r = k$, i.e. when the two feature sets s_1 and s_2 are identical.

2.2 Properties

By leading a thorough study of the literature, we have identified the following set of desirable properties for a stability measure:

1. **Limits.** The measure should be bounded by values that do not depend on the number of features and the cardinality of the feature sets should reach its maximal value when the feature sets are identical.
2. **Monotonicity.** The measure should be an increasing function of the similarity of the feature sets.
3. **Correction for chance.** This property allows to compare stability of FSAs selecting a different amount of features. Positive values will be interpreted as being more stable than an FSA selecting features at random.
4. **Unconstrained on cardinality.** We would like a stability measure to be able to deal with feature sets of different cardinalities.
5. **Symmetry.** We would like the stability measure to be symmetrical, so that its value does not depend on the order on which the feature sets are taken.
6. **Redundancy awareness.** As features can be redundant, we would like a stability measure to reflect on the true amount of redundant information between the feature sets.

Properties 1 to 3 were the ones identified by Kuncheva [8] and Properties 4 to 6 are the ones that we have identified by looking at the measures proposed later on. Table 1 gives us the properties of the most commonly used existing stability measures for FSAs returning a feature set.

Table 1. Properties of stability measures for FSAs outputting a feature set.

	1	2	3	4	5	6
$\overline{sim}_J$ (Dunne et al. [4])	✓	✓		✓	✓	
$\overline{sim}_H$ (Kalousis et al. [6])	✓	✓		✓	✓	
$\overline{sim}_M$ (Yu et al. [13])	✓	✓		✓	✓	✓
$\overline{sim}_K$ (Kuncheva [8])	✓	✓	✓		✓	
$\overline{sim}_L$ (Lustgarten et al. [9])		✓	✓	✓	✓	
$\overline{sim}_W$ (Wald et al. [12])		✓	✓	✓	✓	
Average $nPOG$ (Zhang et al. [14])		✓	✓	✓		
Average $nPOGR$ (Zhang et al. [14])		✓	✓	✓		✓
CW_{rel} (Somol and Novovičová [11])	✓	✓		✓	✓	
γ_k (Krízek et al. [7])	✓	✓			✓	

The focus of this paper is on the stability measures having the important property of correction for chance. Even though the stability measure CW_{rel} (introduced Somol and Novovičová [11]) does not explicitly yield the property of correction for chance, using Theorem 1, we can point out that CW_{rel} asymptotically holds this property when a constant number of features is selected.

Theorem 1. *For a sequence $\mathcal{A}$ containing feature sets of constant cardinality, the stability measure $CW_{rel}(\mathcal{A})$ is asymptotically equivalent to $\overline{sim_K}(\mathcal{A})$ as the number of feature sets approaches infinity.*[1]

All the five stability measures having the property of correction for chance (cf Table 1) are either taken as the average pairwise similarities between the feature sets (as in $\overline{sim_K}$, $\overline{sim_L}$, $\overline{sim_W}$) or as the average similarity between disjoint feature sets pairs (for stability measures using $nPOG$ and $nPOGR$). The $nPOGR$ similarity measure is a generalization of the $nPOG$ measure that attempts to take into account linear feature redundancies (which is not in the scope of this paper). The similarity measures sim_L, sim_W and $nPOG$ are all variants of Kuncheva's similarity measure sim_K for feature sets of varying cardinalities.

3   Extensions of Kuncheva's Similarity Measure

3.1   Definitions

There are three similarity measures extending Kuncheva's similarity measure sim_K for feature sets s_1 and s_2 of different cardinalities (respectively k_1 and k_2). In this situation, the value of the expected size of the intersection for randomly drawn feature sets becomes $\mathbb{E}[r] = \frac{k_1 k_2}{n}$ [9]. The three measures are of the same general form as Kuncheva's measure sim_K, as they keep the numerator equal to $r - \mathbb{E}[r]$. In order to make these values comparable in different settings (i.e. for different values of k_1, k_2 and n), the value of $r - \mathbb{E}[r]$ needs to be rescaled. The three similarity measure extending sim_K are three variants of this and they only differ in the way the numerator $r - \mathbb{E}[r]$ is rescaled. Note that in all these expressions, the only variable is the size of the intersection r and that all other terms are constants only depending on k_1, k_2 and n. Lustgarten et al. [9] proposes dividing the value of the numerator by $r - \mathbb{E}[r]$ by its range (i.e. by its maximal value minus its minimal value for a given k_1, k_2 and n):

$$sim_L(s_1, s_2) = \frac{r - \mathbb{E}[r]}{\max(r - \mathbb{E}[r]) - \min(r - \mathbb{E}[r])}. \tag{5}$$

As $\mathbb{E}[r]$ is a constant only depending on k_1, k_2 and n, $r - \mathbb{E}[r]$ is a linear function of r and hence the above equation becomes:

$$sim_L(s_1, s_2) = \frac{r - \mathbb{E}[r]}{(\max(r) - \mathbb{E}[r]) - (\min(r) - \mathbb{E}[r])} = \frac{r - \mathbb{E}[r]}{\max(r) - \min(r)}, \tag{6}$$

where $max(r)$ and $min(r)$ are respectively the maximal and the minimal possible values of the size of the intersection r given k_1, k_2 and n. Intuitively, we can see that the minimal size of the intersection between two feature sets is not 0. Indeed, imagine we have a set containing $k_1 = 2$ features, another set containing $k_2 = 3$ features and that we have $n = 4$ features to select from in total. In this setting,

---

[1] Proofs of the theorems available at http://www.cs.man.ac.uk/~nogueirs.

these two sets cannot be disjoint. It can be shown that the minimal possible value of r is equal to $min(r) = max(0, k_1 + k_2 - n)$. Similarly the maximal value of r is reached when one set is a proper subset of the other. Therefore, the maximal value or r is equal to $max(r) = min(k_1, k_2)$. Lustgarten's measure sim_L can therefore be rewritten as follows:

$$sim_L(s_1, s_2) = \frac{r - \mathbb{E}[r]}{\max(0, k_1 + k_2 - n) - \min(k_1, k_2)}. \tag{7}$$

It can be shown that this rescaling procedure ensures a value of sim_L in the interval $[-1, 1]$. Nevertheless, we will see through a set of examples that this procedure does not satisfy all the desirable properties in this context.

In a similar way, Wald et al. [12] proposes rescaling the numerator by dividing it by its maximal value:

$$sim_W(s_1, s_2) = \frac{r - \mathbb{E}[r]}{\max(r - \mathbb{E}[r])} = \frac{r - \mathbb{E}[r]}{\max(r) - \mathbb{E}[r]} = \frac{r - \frac{k_1 k_2}{n}}{\min(k_1, k_2) - \frac{k_1 k_2}{n}}. \tag{8}$$

By dividing the numerator by its maximal value, we are ensured that sim_W will always be less than or equal to 1. Nevertheless, as the numerator can take negative values, dividing it by the maximal value will not guarantee lower bounds that do not depend on the constants k_1, k_2 and n. In fact, it can be shown that for a given n, the minimum of sim_W is $1 - n$ (and is reached when $k_1 = n - 1$ and $k_2 = 1$ or vice versa). We will illustrate the importance of this with an example in the next Section. In the measure $nPOG$, Zhang et al. [14] divide the numerator either by $k_1 - \mathbb{E}[r]$ if s_1 is given as the first argument or by $k_2 - \mathbb{E}[r]$ otherwise; making the resulting similarity measure non-symmetrical (i.e. $nPOG(s_1, s_2) \neq nPOG(s_2, s_1)$):

$$nPOG(s_1, s_2) = \frac{r - \mathbb{E}[r]}{k_1 - \mathbb{E}[r]} = \frac{r - \frac{k_1 k_2}{n}}{k_1 - \frac{k_1 k_2}{n}}. \tag{9}$$

The non-symmetry of this measure can be problematic, as we will illustrate it in the next Section. Also, one can notice that when the set of smaller cardinality is given as the first argument, $nPOG$ is equal to the sim_W measure, hence inheriting its weaknesses.

3.2   Toy Examples Illustrating the Weaknesses of the Measures

To illustrate some of the missing properties of the similarity measures, we provide four toy examples.

Example 1: Accounting for Systematic Bias in Chosen Set Size. Imagine that there are 10 features to choose from. Procedure F_1 chooses 7 features *deterministically*, i.e. no matter what the variation in data is, the same 7 features are returned. Intuitively, the "stability" of this procedure is maximal. It has

zero variation in its choice of feature set. Lustgarten's measure returns a value of $sim_L = 0.7$, whilst *all other measures* return 1. This is somewhat strange, and undesirable, as we have no way to know if F_1 is deterministic from the value of sim_L. Furthermore, imagine procedure F_2, which picks 4 features, again deterministically. Lustgarten's measure now returns 0.6, which makes procedures F_1 and F_2 not comparable in terms of stability. This example highlights the need for a similarity measure that returns its maximal value as long as the feature sets are identical, as stated in Property 1 of Sect. 2.

Example 2: Accounting for the Set Size Variations. Imagine our same set of 10 features as above. Half of the time, procedure F returns features 1 to 8, and the other half of the time it returns features 1 and 2, i.e. a proper subset. In this situation, Wald's stability measure returns a maximal 1, whilst clearly there is variation in the choice of the subset size. In fact, using Wald's measure, the similarity between two feature sets will always have its maximal value of 1 as long as one of the two sets is a proper subset of the other. For the other two similarity measures, this is not the case, even though the similarities do not decrease proportionally to the distance between the feature sets cardinalities.

Example 3: Invariance to Feature Set Permutations. This example is the same as the previous one where the order of the feature sets has been permuted. Because of the non-symmetry of $nPOG$, the stability value returned by this measure might not be the same as the one calculated in the previous example. This example showcases the need for a symmetrical similarity measure as stated in Property 5 of Sect. 2.

Example 4: Bounded by Constants. Having minimal values that increase linearly with n can lead to negative values of much larger amplitude than the maximum value which is equal to 1. If we have $n = 100$ features in total, the minimal value of $nPOG$ and of sim_W is equal to -99 while its maximal value is 1. When calculating the average of the similarities, such large negative values can strongly bias the resulting stability. We can illustrate this with a simple example. Imagine that a feature selection procedure selects 9 times features 1 to 8 in feature sets $s_1, s_2, ...s_9$ and that features 9 and 10 are selected in a set s_{10}. When averaging over all possible pairs of similarities, the stability value of $\overline{sim_W}$ is 0 which corresponds to the stability value of an FSA drawing 10 feature sets at random, even though 9 out of the 10 feature sets considered were identical. Another issue with minimal values depending on n is that the minimal values will be different for two different values of n. In practice, this does not allow us to compare the stability of an FSA on two different datasets for instance. This shows the need for a stability measure to be bounded by constants as stated in Property 1 of Sect. 2.

4   A New Similarity Measure

In the light of the previous observations, we propose a new similarity measure sim_N of the same general form that will rescale the numerator $r - \mathbb{E}[r]$ so that its value belongs to $[-1, 1]$. As the numerator $r - \mathbb{E}[r]$ can take both negative and positive values, one way to do so is to divide it by its maximal absolute value as follows:

$$sim_N(s_1, s_2) = \frac{r - \mathbb{E}[r]}{max\left(|r - \mathbb{E}[r]|\right)}. \tag{10}$$

Both Kuncheva's and Wald's similarity measures sim_K and sim_W rescale the numerator by dividing it by its maximal value. Then, we can wonder why Kuncheva's similarity measure sim_K (defined only for $k_1 = k_2$) belongs to $[-1, 1]$ whereas Wald's measure (defined for distinct values of k_1 and k_2) does not. In fact, it can be shown that when $k_1 = k_2$, the maximal absolute value of the numerator is equal to its maximal value, so that Kuncheva's measure can be rewritten as in Theorem 2. This also proves that our proposed measure sim_N is a true generalization of Kuncheva's index as they have the same formal expression.

Theorem 2. *Kuncheva's similarity between two feature sets s_1 and s_2 of same cardinality can be rewritten as follows:*

$$sim_K(s_1, s_2) = \frac{r - \mathbb{E}[r]}{max(|r - \mathbb{E}[r]|)}. \tag{11}$$

The maximal absolute value of a term is equal to the maximum between the opposite of its minimum and its maximum. Therefore, sim_N can be rewritten as follows:

$$sim_N(s_1, s_2) = \frac{r - \mathbb{E}[r]}{max\left[-min(r - \mathbb{E}[r]); max(r - \mathbb{E}[r])\right]}. \tag{12}$$

The only variable in sim_N is the size of the intersection r and all other terms only depend on k_1, k_2 and n. Therefore, $min(r - \mathbb{E}[r]) = min(r) - \mathbb{E}[r]$ and $max(r - \mathbb{E}[r]) = max(r) - \mathbb{E}[r]$, which gives us the following expression for sim_N:

$$sim_N(s_1, s_2) = \frac{r - \mathbb{E}[r]}{max\left[-min(r) + \mathbb{E}[r]; max(r) - \mathbb{E}[r]\right]}. \tag{13}$$

As explained in Sect. 3.1, the minimal value of r is $min(r) = max(0, k_1 + k_2 - n)$ and its maximal value is $max(r) = min(k_1, k_2)$. Therefore, we have that:

$$sim_N(s_1, s_2) = \frac{r - \mathbb{E}[r]}{max\left[-max(0, k_1 + k_2 - n) + \mathbb{E}[r]; min(k_1, k_2) - \mathbb{E}[r]\right]}$$
$$= \frac{r - \frac{k_1 k_2}{n}}{max\left[-max(0, k_1 + k_2 - n) + \frac{k_1 k_2}{n}; min(k_1, k_2) - \frac{k_1 k_2}{n}\right]}. \tag{14}$$

The resulting stability measure $\overline{sim_N}$ is then taken as the average pairwise similarities as in Eq. 1. Let us now look at the properties of the new measure sim_N.

As explained previously, this measure is a proper generalization of Kuncheva's measure sim_K as it matches its value for $k_1 = k_2 = k$. By construction, this measure will be bounded by the constants -1 and 1 and reach its maximal value of 1 when the two feature sets are identical. Hence it has the Property 1 of Table 1. As outlined in the toy examples of Sect. 3.2, this allows the comparison of stability values for algorithms returning different number of features and for different values of n. It also has the Property 2 of monotonicity (as it is an increasing function of the size of the intersection r between two feature sets) and the Property 3 of correction for chance. It is invariant to feature set permutations (as it is symmetric). Figure 1 shows the maximum and minimal values of the measure in different settings. As we can see, even though this measure accounts for some of the set size variations, the maximum value is not proportional to the distance between the two subset sizes k_1 and k_2. Finally, this measure is the only one having Properties 1 to 5. As Kuncheva's similarity measure sim_K (as well as the measures sim_L, sim_W and $nPOG$), the expression of this similarity measure only holds for values of k_1 and k_2 in $\{1, ..., n-1\}$. For completeness, we will also set the values of sim_N to 0 when k_1 or k_2 is equal to 0 or n.
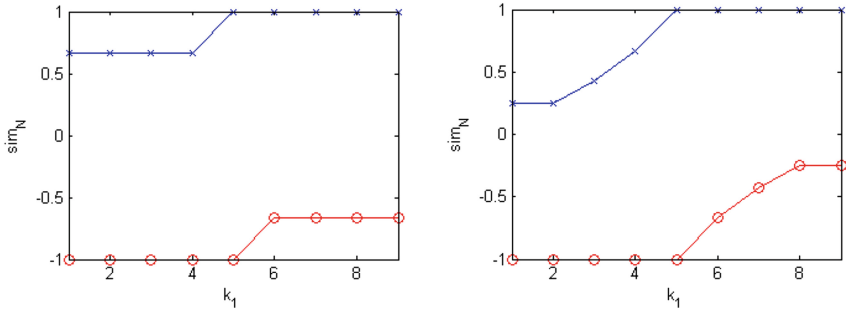

Fig. 1. Maximum and minimum of sim_N against k_1 for $k_2 = 6$ (LEFT) and $k_2 = 8$ (RIGHT) when $n = 10$.

5   Application to Feature Selection by Random Forests

To illustrate the utility of the proposed measure, we used random forests [2] as a feature selection procedure where a feature is selected when it is used in at least a percentage p of the trees. We built random forests of 100 decision trees using the mutual information as a splitting criterion. Each decision tree is built on a bootstrap sample of the given dataset. At each splitting point, the decision tree was given the choice between $\lfloor \sqrt{n} \rfloor$ features selected at random. As p is effectively a regularization parameter on the number of features selected, we tuned p for the different datasets so that only a certain proportion of the features is selected. In order to model the data perturbations, either bootstrap samples or random sub-samples can be taken [5].

Table 2. Parameters of 4 datasets, where p is the regularization parameter and where the average number of features selected using parameter p is given along with its standard deviation.

	Num. ex.	Num. classes	p	Num. feat.	Av. num. feat. selected
Wine	178	3	0.5	13	9.8 ± 0.93
Parkinsons	195	2	0.5	22	11 ± 1.7
Breast	569	2	0.5	30	15 ± 1.3
Sonar	208	2	0.25	60	42 ± 2.4

(a) wine
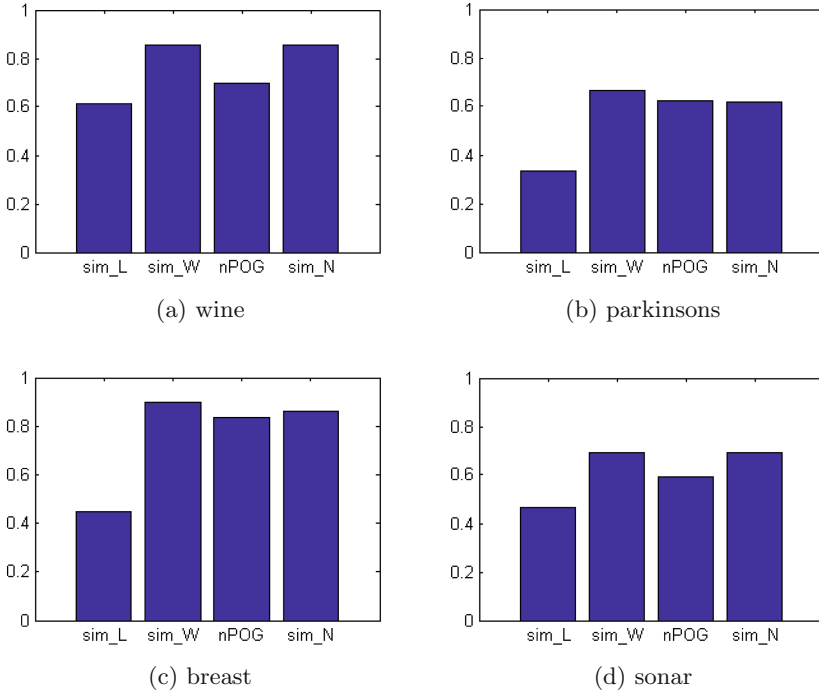
(b) parkinsons

(c) breast

(d) sonar

Fig. 2. Stability values on 4 datasets using the different similarity values.

Here we used $K = 100$ random sub-samples without replacement of the datasets containing 90 % of the total amount of examples [8]. So we built K random forests on each one of these samples and calculated the stability of the sequence of the K feature sets obtained. We used 4 datasets of the UCI repository, for which the properties, the values chosen for p and the average number of features selected are given in Table 2. Figure 2 gives us the stability values when using the different similarity measures. We observed that on all the datasets, the lowest stability value is obtained when using Lustgarten's similarity measure sim_L. This probably comes from the fact that sim_L does not always reach its maximal value when two feature sets are identical and its maximal value

depends on the size of the feature sets selected (as observed in Toy Example 1 of Sect. 3.2). The stability value when using $nPOG$ seems closer to the value of the ones using sim_N and sim_W on the parkinsons and on the breast datasets than in the other two datasets. As we have seen in Toy Example 3, the value of $nPOG$ changes when we permute the feature sets, which makes it difficult to interpret. On the four datasets, the stability values obtained using sim_W and sim_N are close to each other. This can be explained by the fact that in some situations (i.e. for certain values of k_1, k_2 and n), the value sim_N will be equal to the one of sim_W. Indeed, when we take a pair of feature sets s_1 and s_2, if we have k_1, k_2 and n such that $-min(r) + \mathbb{E}[r] \leq max(r) - \mathbb{E}[r]$, the denominator of the two similarity measures becomes the same and in that case $sim_W(s_1, s_2) = sim_N(s_1, s_2)$. In other words, in the feature sets returned by this procedure, only a small proportion of pairs of feature sets do not satisfy this. We have seen in Toy Example 4 that the minimal value of sim_W decreases with n and this could strongly bias the resulting stability value in some cases. This situation happens when the feature sets are very dissimilar in both terms of cardinality and of features selected. In the four datasets, we can observe that this is not the case as the standard deviations of the number of features selected by the random forests are much smaller than the total number of features.

6   Conclusion

Through a thorough study of the literature, we identified a set of desirable properties for stability measures dealing with feature selection procedures that return feature sets. After leading a comparative study on the measures that have the property of correction for chance, a generalization of Kuncheva's index is proposed for feature selection algorithms that do not return feature sets of constant cardinality. This new measure has all the desired properties except that it does not take into account possible redundancy between features, which could be the focus of future work. We illustrate a possible application of this measure in the context of ensemble-based feature selection and exhibit the differences obtained in the stability values using the different measures.

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**, 392–398 (2010)
2. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
3. Ditzler, G., Polikar, R., Rosen, G.: A bootstrap based neyman-pearson test for identifying variable importance. IEEE Trans. Neural Netw. Learn. Syst. **26**, 880–886 (2014)

4. Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical report, Journal of Machine Learning Research (2002)
5. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Comput. Biol. Chem. **34**, 215–225 (2010)
6. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. **12**, 95–116 (2007)
7. Křížek, P., Kittler, J., Hlaváč, V.: Improving stability of feature selection methods. In: Kropatsch, W.G., Kampel, M., Hanbury, A. (eds.) CAIP 2007. LNCS, vol. 4673, pp. 929–936. Springer, Heidelberg (2007)
8. Kuncheva, L.I.: A stability index for feature selection. In: Artificial Intelligence and Applications (2007)
9. Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S.: Measuring stability of feature selection in biomedical datasets. In: Proceedings of the AMIA Annual Symposium (2009)
10. Saeys, Y., Abeel, T., Van de 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)
11. Somol, P., Novovičová, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 1921–1939 (2010)
12. Wald, R., Khoshgoftaar, T.M., Napolitano, A.: Stability of filter- and wrapper-based feature subset selection. In: International Conference on Tools with Artificial Intelligence. IEEE Computer Society (2013)
13. Yu, L., Ding, C.H.Q., Loscalzo, S.: Stable feature selection via dense feature groups. In: KDD (2008)
14. Zhang, M., Zhang, L., Zou, J., Yao, C., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, C., Guo, Z.: Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinformatics **25**, 1662–1668 (2009)