Evaluating Stability of Feature Selectors: Adjusted Measures Considering Feature Correlations

1st Aiguo Wang
School of Electronic Information Engineering
Foshan University
Foshan, China
wangaiguo2546@163.com

2nd Jingyu Yan
School of Electronic Information Engineering
Foshan University
Foshan, China
yanjingyu163@163.com

3th Zhongyu Luo
School of Electronic Information Engineering
Foshan University
Foshan, China
zhongyu.jojo@outlook.com

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995507

***Abstract*—Besides the aim of identifying a subset of useful features, the stability of feature selection algorithms is also a critical topic in increasing the confidence of selected features, where an objective stability measure with required properties is expected. To this end, we herein propose a new stability measure sim_{nJR} that considers the feature correlations and possesses the desirable properties. Specifically, we first utilize the Pearson correlation coefficients and the false discovery rate control procedure to identify significantly correlated feature pairs that are not shared in two feature sets. A normalization step is then conducted to reduce the effects of the size of feature sets and of general feature correlations in the dataset. Finally, we consider two commonly used feature selectors (i.e., relief and mRMR) and conduct comparative experiments on several datasets under different hyperparameter values and the variation of train sets. Results show its effectiveness.**

Keywords—feature selection, stability, similarity, Pearson correlation coefficients

I. INTRODUCTION

Feature selection, as an important preprocessing technique, is widely used in a variety of tasks such as computer vision, bioinformatics, and natural language processing. Accordingly, researchers have proposed a large number of feature selection algorithms that can be broadly categorized into filter, wrapper, and embedded methods towards a subset of discriminant features [1]. The use of feature selection helps improve the prediction accuracy, facilitate the interpretability, provide insights into knowledge discovery, and reduce the cost of data collection [2]. On the other hand, the stability of feature selector is also an important aspect in better uncovering the underlying data mechanism, since correlated features generally exist in the dataset and can produce multiple equally discriminant feature sets [3]. Stability indicates that a good feature selector should be robust to the perturbation of training data and to its different hyperparameter values (if existing) [4]. In contrast, an instable feature selector would greatly lower the confidence of selected features and prevent users from performing further analysis, especially in the biomedical and bioinformatics fields [3].

The sources of feature selection instability come from the *data level* (e.g., correlated features exist in the dataset) and *algorithm level* (e.g., a feature selector is sensitive to dataset variation) [5]. Current studies about feature selection stability mainly concerns how to effectively improve the stability and how to accurately measure the stability of feature selectors [6]. As for the latter, we can categorize existing stability measures into frequency- and similarity-based measures according to the methodology. Specifically, given a collection of feature sets $S = \{s_1, s_2, \cdots, s_Q\}$, where $s_i (1 \le i \le Q)$ is chosen from the original feature set F via a feature selection algorithm, frequency-based measures consider the frequency of selection of each feature over the Q sets [7]. For example, the weighted consistency metric uses Eq. (1) to calculate the stability score.

$$CW(S) = \sum_{f \in union(S)} \frac{C_f}{\sum_{f \in union(S)} C_f} \times \frac{C_f - C_{\min}}{C_{\max} - C_{\min}} \tag{1}$$

, where $C_f = \sum_{i=1}^{Q} I(f \in s_i)$ and C_{min} (C_{max}) is the minimum (maximum) occurrences of $f \in union(S) = \cup(s_1, s_2, \mathrm{L}\ , s_Q)$.

Similarity-based measures work in a local scheme and use the average pairwise similarity between each pair of feature sets in S to measure the stability [8, 9], as shown in Eq. (2).

$$\varphi(S) = \frac{2}{Q(Q-1)} \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} sim(s_i, s_j) \tag{2}$$

, where $sim(s_i, s_j)$ denotes the similarity between s_i and s_j. Obviously, the key of such methods is how to evaluate the similarity between two sets. Accordingly, researchers have proposed a number of metrics such as Lustgarten's measure (sim_L), normalized percentage of overlapping genes-related (sim_{nPOGR}), and Jaccard index (sim_J) shown in Eq. (3).

$$sim_J = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \tag{3}$$

To help users to identify the appropriate stability measure, researchers have summarized the following six properties that a stability measure is expected to possess.

(1) *Bounds*. A stability measure should be unconstrained on cardinality of feature sets and be upper/lower bounded by constants, and the measure achieves the maximum when *if-and-only-if* the feature sets are identical.

(2) *Symmetry*. A stability measure is irrelevant to the order of feature sets, i.e., $sim(s_i, s_j) = sim(s_j, s_i)$.

(3) *Monotonicity*. The more similar the feature sets are, the higher stability score the measure is.

(4) *Fully defined*. A stability measure should be defined for feature sets of different cardinalities, indicating s_i and s_j can contain different number of features.

(5) *Correction for chance*. The expectation of a stability measure is a constant when the feature selection is random.

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515011499).

(6) *Redundancy awareness*. Since correlated features exist in the dataset, a stability measure should consider redundant information among features to adjust the stability score.

With those properties, we can see that Jaccard index fails to satisfy the fifth and sixth properties, although it has the advantages of simplicity and intuition. To this end, we herein propose an adjusted measure that takes into account feature correlations and correction for chance. Specifically, we use the Pearson correlation coefficients and false discovery rate to judge whether feature pairs are significantly correlated and then conduct a normalization step to reduce the effects of general feature correlations. Our main contributions include the follows. (1) We propose an improved version of Jaccard index that considers feature correlations. This helps us better evaluate the stability of a feature selector. Table I summarizes the comparisons between our work and related work, which shows that the proposed sim_{nJR} owns the desirable properties of a stability measure. (2) We utilize two commonly used feature selection algorithms and do comparative experiments on public datasets. Results show its effectiveness.

Table I. Property Comparisons of Different Stability Measures

	1	2	3	4	5	6
sim_K	✓	✓	✓		✓	
sim_J	✓	✓	✓	✓		
sim_{nJ}	✓	✓	✓	✓	✓	
sim_{JR}	✓	✓	✓	✓		✓
sim_{nJR}	✓	✓	✓	✓	✓	✓

II. The Proposed Stability Measure

If $sim(s_i, s_j)$ is given for $S = \{s_1, s_2, \cdots, s_Q\}$, we can get the stability score using Eq. (1). Considering that correlated features exist in the dataset [10, 11], the metric sim_{JR} in Eq. (4) is used to measure the consistency between s_i and s_j.

$$sim_{JR} = \frac{\left|s_i \cap s_j\right| + R_{ij}}{\left|s_i \cup s_j\right|} \tag{4}$$

, where R_{ij} is the number of features that are not shared in s_i and s_j but significantly correlated with at least one feature in s_i or s_j. Specifically, for $f \in s_i$ (s_j) that is not shared in s_i and s_j, we evaluate whether at least one feature $g \in s_j$ (s_i) significantly correlated with *f* exists. To decide whether the correlation *r* between *f* and *g* is significant, we can of course use a threshold θ. If $r \geq \theta$, the correlation is significant. We here adopt a statistical test method. We first permute the values of each feature in *F*, calculate the Pearson correlation coefficients for all feature pairs, and obtain the *P*-value that equals the percentage of permutated correlations greater than *r*. We then use the false discovery rate control procedure to obtain the significantly correlated feature pairs [3].

Afterwards, we use Eq. (5) to normalize the effects of the size of feature sets and mitigate the effects of general feature correlations in the dataset.

$$sim_{nJR} = \frac{sim_{JR} - E(sim_{JR})}{1 - E(sim_{JR})} \tag{5}$$

, where $E(sim_{JR})$ is the expectation of sim_{JR} and is estimated in this study by the expectation of the scores for 10000 pairs of feature sets randomly chosen from *F* (with lengths $|s_i|$ and $|s_j|$). Consequently, sim_{nJR} would be more appropriate for measuring stability of feature sets. Obviously, sim_{nJR} equals 1 if $sim_{JR} = 1$ and $E(sim_{JR}) < 1$. We define $sim_{nJR} = 0$, if $sim_{JR} \leq E(sim_{JR}) < 1$.

Likewise, we could normalize the sim_J using eq. (6).

$$sim_{nJ} = \frac{sim_J - E(sim_J)}{1 - E(sim_J)} \tag{6}$$

, where $E(sim_J)$ is the expectation of sim_J and estimated by the mean of scores for 10000 pairs of feature sets randomly chosen from *F* (with lengths $|s_i|$ and $|s_j|$). If $sim_J = 1$ and $E(sim_J) < 1$, $sim_{nJ} = 1$; $sim_{nJ} = 0$, if $sim_J \leq E(sim_J) < 1$.

III. Experimental Results and Analysis

A. Experimental Setup

To evaluate the proposed stability measures, we conduct extensive comparative experiments on five UCI datasets that cover both binary and multi-classes cases, as shown in Table II. As for feature selection algorithms, we include reliefF and min-Redundancy Max-Relevance (MRMR) as a comparison. For the stability measures, we use sim_J and its competitors (i.e., sim_{nJ}, sim_{JR}, and sim_{nJR}) to measure the robustness of a feature selection algorithm to its different hyperparameter values and to the variation of training set. Besides evaluating the stability of a feature selector, we evaluate its predictive power. Hence, *k*-nearest-neighbors (KNN) and support vector machine with linear kernel (SVM) are used [12].

Table II. Experimental Datasets

ID	Dataset	#samples	#features	#classes
1	*control*	600	60	6
2	*heart*	270	13	2
3	*solar*	323	12	6
4	*vote*	435	16	2
5	*zoo*	101	16	7

B. Stability under Different Hyperparameter Values

We in this section evaluate the stability of reliefF regarding its hyperparameter *k* (i.e., the number of neighbors used to decide the importance of a feature). We first use two different values of *k* (i.e., 5 and 10) to obtain two feature subsets and then calculate the stability scores. Table III presents the results of the top-ten selected features, from which we observe that reliefF is sensitive to its hyperparameter values and that the measures involving feature correlations generally have higher scores (except for the case of sim_{nJR} on *heart*, which is mainly due to the high general feature correlations.

C. Stability under the Variation of Datasets

We in this section evaluate the stability of feature selectors under the variation of training sets. Specifically, a 5-fold cross validation scheme is utilized to partition the original dataset into five folds, where each one of the five folds is used as a test set and the remaining folds are used as the training data. Notably, feature selection is only conducted on the training data, and hence this procedure obtains five feature subsets. Since reliefF and mRMR belong to feature ranking algorithms, two different thresholds (i.e., five and ten) are used to get the

Table III. Stabilty of ReliefF with Different Hyperparameter Values

Dataset	sim_J	sim_{nJ}	sim_{JR}	sim_{nJR}
control	0.25	0.21	1	1
heart	0.43	0.23	0.86	0.06
solar	0.43	0.21	1	1
vote	0.67	0.58	1	1
zoo	0.43	0.29	1	1

finally selected features. Afterwards, we calculate the stability score using Eq. (1). Experimental results related to reliefF and mRMR are shown in Tables IV and V, respectively.

From Table IV, for sim_J, we can observe that the number of selected features influences the stability scores and that a larger number of selected features generally gets higher scores in the majority of cases. Second, we observe that the inclusion of the correction for chance generally decreases the stability scores. For example, if ten features are selected from *zoo*, sim_J and sim_{nJ} get the stability scores of 0.79 and 0.61, respectively. Third, we can see that the consideration of feature correlations generally obtains higher scores. For example, if ten features are selected from *zoo*, the score of sim_{JR} is 0.98 compared to 0.79 of sim_J, and sim_{nJR} enhances the score of sim_{nJ} from 0.61 to 0.92. This indicates the existence of correlated features. Similar results can be observed in Table V.

Table IV. Stability Scores Using reliefF

Dataset		sim_J	sim_{nJ}	sim_{JR}	sim_{nJR}
control	5	0.42	0.39	1	1
	10	0.69	0.65	1	1
heart	5	0.6	0.46	0.99	0.93
	10	0.71	0.25	0.96	0.5
solar	5	0.74	0.64	1	1
	10	0.89	0.61	1	1
vote	5	0.7	0.63	1	1
	10	0.73	0.49	0.97	0.6
zoo	5	0.8	0.75	0.9	0.74
	10	0.79	0.61	0.98	0.92

Table V. Stability Scores Using mRMR

Dataset		sim_J	sim_{nJ}	sim_{JR}	sim_{nJR}
control	5	0.49	0.46	1	1
	10	0.54	0.50	1	1
heart	5	1	1	1	1
	10	0.93	0.80	1	1
solar	5	1	1	1	1
	10	0.87	0.55	0.99	0.91
vote	5	1	1	1	1
	10	0.89	0.8	1	1
zoo	5	0.7	0.63	0.9	0.75
	10	0.89	0.8	0.98	0.93

D. Stability versus Accuracy

Considering that stability and accuracy are two key aspects in evaluating a feature selection algorithm, we experimentally present their relationships with an aim to direct the choice of feature selectors, where a five-fold cross validation is utilized to generate independent training sets and test sets. Figure 1 shows the results of reliefF and mRMR on *control* when they are combined with KNN and SVM to obtain the accuracy. The X-axis denotes the stability scores of sim_{nJR} and Y-axis refers to prediction accuracy. Obviously, feature selectors located in the upper right corner remain a priority.

IV. Conclusion

Stability is an important aspect in evaluating the power of a feature selector, where the design of a stability measure with desirable properties is of great value. We in this study propose an adjusted measure that considers feature correlations. First, the Pearson correlation coefficients and false discovery rate are used to identify the significantly correlated feature pairs that are not shared in two feature sets. Then, a normalization step is conducted. Finally, we conduct experiments using two feature selectors, and results indicates its effectiveness. For future research, we can apply the metric to high-dimensional datasets such as microarray and RNA-seq data.

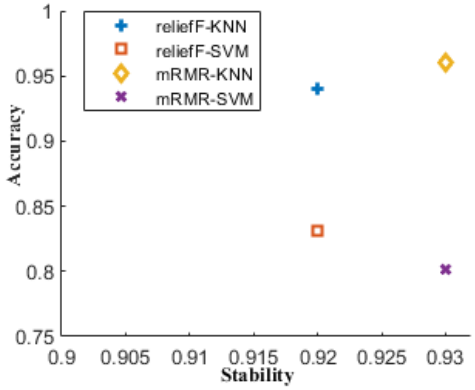

Fig. 1. Stability vs. accuracy of different feature selectors on *control*.

References

[1] G. Brown, A. Pocock, M. J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 27-66, 2012.

[2] A. Wang, N. An, G. Chen, L. Li, and G. Alterovitz. "Accelerating wrapper-based feature selection with K-nearest-neighbor," *Knowledge-Based Systems*, vol. 83, pp. 81-91, 2015.

[3] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, et al. "Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes," *Bioinformatics*, vol. 25, no. 13, pp. 1662-1668, 2009.

[4] S. Nogueira, K. Sechidis, and G. Brown. "On the stability of feature selection algorithms," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6345-6398, 2017.

[5] A. Wang, H. Liu, J. Yang, and G. Chen. "Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data," *Computers in Biology and Medicine*, vol. 142, p. 105208, 2022.

[6] U. Khaire, and R. Dhanalakshmi. "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060-1073, 2019.

[7] P. Somol and J. Novovicova, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1921-1939, 2010.

[8] S. Nogueira, and G. Brown. "Measuring the stability of feature selection," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 442-457, 2016.

[9] K. Sechidis, K. Papangelou, S. Nogueira, J. Weatherall, and G. Brown. "On the stability of feature selection in the presence of feature correlations," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 327-342, 2019.

[10] A. Bommert, and J. Rahnenführer. "Adjusted measures for feature selection stability for data sets with similar features," *International Conference on Machine Learning, Optimization, and Data Science*, pp. 203-214, 2020.

[11] M. Zucknick, S. Richardson, and E. Stronach. "Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, pp. 1-28, 2008.

[12] A. Wang, A. Ning, G. Chen, L. Liu, and G. Alterovitz. "Subtype dependent biomarker identification and tumor classification from gene expression profiles," *Knowledge-Based Systems*, vol. 146, pp. 104-117, 2018.