Abstract—The feature selection is an inevitable part of machine learning techniques in biomedical engineering and bioinformatics. Feature selection methods are used to select the most discriminative features, e.g. for disease classification. Even if there are plenty of feature selection methods the stability of these algorithms is still an open question. Another issue with assessing the stability of feature selection is that there are several stability measures providing different views on stability. Here, we compare well-known stability measures and evaluate their performance on artificial and real data.

I. INTRODUCTION

Machine learning (ML) as a sub-discipline in the field of artificial intelligence has become an eminent part of biomedical research. It resulted in establishment of new research areas such as bioinformatics or health informatics. Machine learning techniques represent tools to improve disease classification, detection of irregularities or increase objectivity of medical decision making. This is confirmed by many recent studies [1], [2], [3] that proved importance of ML techniques and contributed to more precise and objective diagnostic tools [4].

In the past years, dimensionality of biomedical datasets have expanded from tens to thousands of features or variables giving rise to the curse of dimensionality [5]. High dimensionality occurs when the number of features is much higher than the number of instances or subjects in dataset. One of the solutions to cope with this challenging issue is to employ feature selection (FS). The goal of FS is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results [6], [7].

When solving machine learning problems researchers frequently make use of dataset perturbation. Dataset perturbation consists of randomly removing instances from a dataset in order to create one or more reduced datasets [8]. Applying FS on perturbed dataset leads to different subsets of selected features. Robustness of the feature preferences produced by FS relative to variability of perturbed datasets is defined as feature selection stability [9]. Stable FS techniques are strongly preferred since they provide feature subset that can be used even when there is a change in dataset. Recently, there have been several studies that compare stability of FS techniques with respect to variations in dataset [10], [11], [12], [13], [14] or influence of classifier in single feature classifier FS [15]. Since there is no single framework for evaluation of FS stability many different feature selection metrics are available to evaluate stability of FS process. It is hard to decide which FS measure is the best suited for particular problem and there are not many direct comparisons of these metrics.

Here, we compare several known metrics for feature selection stability. We analyse performance of these metrics on artificial case study examples as well as real biomedical datasets. Our aim is to obtain better insight into stability measures and their performance.

The rest of this paper is organized as follows. In the next section we provide description of features selection techniques followed by definition of stability measures that are used to evaluate stability of FS algorithms. Then, numerical results on artificial and real data are provided. Finally conclusions are drawn in the last section.

II. FEATURE SELECTION ALGORITHMS

Feature selection techniques found many applications in bioinformatics for sequence or microarray analysis, mass spectra analysis and text or literature mining.

Based on the way how they combine the feature selection search with the construction of the classification model the FS algorithms can be divided into three groups: filter techniques, wrapper methods and embedded techniques [16]. Filter techniques are independent of classifier and are usually computationally simple and fast. Even if these techniques are rather simple and univariate, several studies indicate that they can perform as good as more complex methods [10], [12]. Wrapper methods involve optimizing a predictor as part of the selection process. This usually comes with significantly higher complexity compared to filter methods. Main advantage of these methods is interaction between feature subset search and model selections. On the other hand, wrapper algorithm have high risk of overfitting. The embedded methods use classifier algorithm to built an optimal subset of features. Similarly to wrapper methods the embedded method is specific to particular learning algorithm.

When repeatedly applying FS on perturbed dataset, the subset of selected features varies. This is illustrated in Fig.1 showing frequency that a specific feature appears during the FS process after applying 500 perturbations selecting 100 features every time. This illustrates relatively high variability at the output of FS.
III. FEATURE SELECTION STABILITY MEASURES

In recent years there has been a number of different stability measures implemented to assess robustness of feature selection techniques. Different stability measures express slightly different aspect of the problem, however there are some properties that are common for group of measures. Following [17] the FS measures can be divided according to several criteria. Measure evaluating overall feature occurrence frequency over the system as a whole are denoted as feature-focused. On the other hand, measures evaluating features with respect to their occurrence in each particular subset in the system are denoted as subset focused. Then, based on the importance given to feature exclusion, we distinguish between selection registering (ignore the information on the stability of feature exclusion) and selection exclusion registering (take into account both the stability of presence and the absence of features in subsets) measures.

Let \( F = \{f_1, \ldots, f_c\} \) be the set of all features of cardinality \(|F| = c\) and \( S = \{S_1, \ldots, S_K\} \) be a system of \( K \) feature subsets, obtained by applying \( K \) times particular FS algorithm on different samplings of data set. Let \( S_{id} \) and \( S_{jd} \) be the subset of features \( S_{id}, S_{jd} \subseteq F \), where \(|S_{id}| = |S_{jd}| = d\).

Then, Kuncheva index \( \kappa(S) \) [18] for a system \( S = \{S_{id}, \ldots, S_{Kd}\} \), for fixed subset size \( d \leq c \) is defined as

\[
\kappa(S) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{|S_{id} \cap S_{jd}| \cdot c - d^2}{d(c - d)}.
\]  

Kalousis [9] suggested stability index \( \Psi(S) \) based on Tanimoto distance, which measure dissimilarity between two subsets:

\[
\Psi(S) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{|S_{id} \cap S_{jd}|}{|S_{id} \cup S_{jd}|}.
\]  

Next, we define Dunne stability index [19] as

\[
D(S) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{|S_{id} \setminus S_{jd}| + |S_{jd} \setminus S_{id}|}{c}.
\]  

Conceptually different measure is weighted consistency [17]. Let \( N = \sum_{i=1}^{K} S_i \) be the total number of occurrences of any feature in \( S \) and \( N_f \) be the number of occurrences of feature \( f \in F \) in system \( S \). Weighted consistency index \( CW \) is defined as follows [17]:

\[
CW(S) = \sum_{f \in F} \frac{N_f}{N} \cdot \frac{N_f - 1}{K - 1}.
\]

We employ relative weighted consistency index \( CW_{rel} \) to suppresses the influence of the sizes of subsets in system on the measure final value [17], [20]. \( CW_{rel} \) is obtained by adjusting \( CW \) on its minimal \( CW_{min} \) and maximal \( CW_{max} \) possible values as

\[
CW_{rel}(S, F) = \frac{CW(S) - CW_{min}(N, K, F)}{CW_{max}(N, K) - CW_{min}(N, K, F)}
\]  

Introducing \( D = N \mod c \) and \( H = N \mod K \) the \( CW_{rel} \) becomes [20]:

\[
CW_{rel}(S, F) = \frac{c(N - D + \sum_{f \in F} N_f(N_f - 1)) - N^2 + D^2}{c(H^2 + K(N - H) - D) - N^2 + D^2}.
\]

Finally, we introduce simple histogram \( H(S) \) based stability measure. \( H(S) \) express the stability as the ratio between average number of feature occurrences of \( T \) features with highest occurrence and average number of feature occurrences of other features. Let \( F_{top} \subseteq F \) contain \( T \) features with highest occurrence \( N_f \) and \( F_{other} = F \setminus F_{top} \) contain all other features with \( N_f \neq 0 \). Here, \( F_0 = \{f_1, \ldots, f_q\} \), where occurrences \( N_{f_1} = \ldots = N_{f_q} = 0 \). Then

\[
H(S) = \frac{1}{1 + \frac{1}{|F_{top}|} \sum_{f \in F_{top}} N_f}.
\]

IV. EXPERIMENTAL RESULTS

A. Results on artificial data

We analyze behavior of FS stability measures under three different scenarios. Influence of randomness in FS process, subset size and robustness is evaluated and compared.

Firstly, we follow and construct example from [18]. Assume scenario where number of all features \( c = 10 \) and \( K = 2 \) runs of FS algorithm were performed to obtain subset of selected features. Fig. 2 shows the values of stability measures as the function of subset size \( d \). First \( d \) features of \( S_1 \) and \( S_2 \) are included in subset of selected features. The \( S_1 \) and \( S_2 \) are as follows:

\[
S_1 = \{x_9, x_7, x_2, x_1, x_3, x_{10}, x_8, x_4, x_5, x_6\}
\]

\[
S_2 = \{x_3, x_7, x_9, x_{10}, x_2, x_4, x_8, x_6, x_1, x_5\}
\]

As we can see all stability measures correctly identified decrease in stability at \( d = 4 \). Moreover, we can note that there is different behavior of stability measures with increasing subset size for \( d > 5 \). Value of one group of measures \((CW(S), D(S), \psi(S), H(S))\) is increasing due to growing subset size. Since more features from dataset are included the stability by chance contribute to result. The other
two measures ($CW_{rel}(S), \kappa(S)$) have implemented correction for chance and are more robust.

In the next scenario, we explore even more argument for correction by chance. As proposed in [18] we generate ten independent random sequences. Here, only 3 measures provide stability index close to zero: $CW_{rel}(S), \kappa(S)$ and $H(S)$. The value of other measure increase with growing subset size, indicating misleading behavior of these measures.

Now, let us evaluate properties of the stability measures under the condition of increased dimensionality of dataset. We generated 100 random sequences of size $c$ varying from 51 to 30000. The subset size was constant $d = 50$ for all cardinalities. The results are provided in Fig. 4.

**B. Results on real data**

To validate the results also on real databases we evaluate feature selection stability on two biomedical datasets: Golub [21] and PaHaw [22] database. Golub [21] is high dimensional database containing 72 samples with 7129 features. There are two classes of samples 36 from acute lymphoblastic leukemia (ALL) patients and 37 from acute myelogenous leukemia (AML). PaHaw database is significantly smaller database containing handwriting samples from Parkinson’s disease patients (37) and healthy controls (38). The database contains 204 features.

As a feature selection method we had decided to use Univariate FS method based on ANOVA. FS rank all features in terms of a relevance as measured by score provided by ANOVA F-value. The advantage of univariate FS is its simplicity and low computational time. Even if rather simple it has been shown that under some conditions univariate FS performs better than much more complex wrapper or embedded methods [10].

The values of stability measures for Golub database are depicted in Fig. 5. Similarly, stability of feature selection for PaHaw database is presented in Fig. 6. As can be seen from Fig. 5 and Fig. 6 Kuncheva index $\kappa(S)$ and $CW_{rel}$ provide very similar concise results. These results indicate that both measures are concise and are robust to changes in subset size. In contrary, disadvantage of histogram based measure $H(S)$ is that it is influenced by high dimensionality of dataset. This is true also for the case of artificial data example in Fig. 4. Dunne stability measure $D(S)$ shows unsatisfactory behavior, providing values close to 1 for large datasets.

**V. CONCLUSION**

We compared several frequently used feature selection stability measures. Behavior of these measures was analyzed in three different types of artificial data evaluating influence of data dimensionality and size of selected subset. Additionally, stability measures were evaluated by application of univariate features selection technique on two biomedical datasets. Based on presented analysis the Kuncheva index $\kappa(S)$ and weighted consistency $CW_{rel}$ appears to be most concise and precise stability measures.
The described research was performed in laboratories supported by the SIX project; the registration number CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation. Peter Drotár was supported by the project CZ.1.07/2.3.00/30.0005 of Brno University of Technology.

ACKNOWLEDGMENT

REFERENCES


