Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Stability of feature selection algorithm: A review

Utkarsh Mahadeo Khaire [a,*], R. Dhanalakshmi [b]

[a] Department of Computer Science and Engineering, National Institute of Technology Nagaland, Dimapur, India
[b] Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal, India

ARTICLE INFO

Article history:
Received 29 March 2019
Revised 10 June 2019
Accepted 24 June 2019
Available online xxxx

Keywords:
Feature selection
Knowledge discovery
Stability
Robustness
Instability
Perturbation

ABSTRACT

Feature selection technique is a knowledge discovery tool which provides an understanding of the problem through the analysis of the most relevant features. Feature selection aims at building better classifier by listing significant features which also helps in reducing computational overload. Due to existing high throughput technologies and their recent advancements are resulting in high dimensional data due to which feature selection is being treated as handy and mandatory in such datasets. This actually questions the interpretability and stability of traditional feature selection algorithms. The high correlation in features frequently produces multiple equally optimal signatures, which makes traditional feature selection method unstable and thus leading to instability which reduces the confidence of selected features. Stability is the robustness of the feature preferences it produces to perturbation of training samples. Stability indicates the reproducibility power of the feature selection method. High stability of the feature selection algorithm is equally important as the high classification accuracy when evaluating feature selection performance. In this paper, we provide an overview of feature selection techniques and instability of the feature selection algorithm. We also present some of the solutions which can handle the different source of instability.

Contents

* Corresponding author.
E-mail address: utkarshkhaire@gmail.com (U.M. Khaire).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2019.06.012
1319-1578/

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

1. Introduction

Recent advancement in high-throughput technologies such as Next-generation sequencing (NGS), Microarray, Mass spectrometry (MS), etc. open the new gateways for researchers to identify the genetic cause of diseases (Mohammadi et al., 2016; Taylor et al., 2015). Radiomics is an emerging technology where a medical image provides crucial information regarding tumour physiology (Parmar et al., 2015). These high-throughput representations suffers from curse of dimensionality therefore it require a proper computational method to extract knowledge from them (Hinrichs et al., 2019). Microarray data contain many heterogeneity factors as it contains an expression of every possible gene in the genome. It scientifically proved that the genes which are responsible for some biological process are interrelated to each other and some genes are activators or inhibitors of others (Perthame et al., 2016). In high-dimensional data, such as microrray datasets irrelevant features can interfere with the true features, which in turn introduces heterogeneity in the data and generate dependence across the features. Statistical analysis loses its importance in case of dependent features. So, we have to select features that play a vital role in estimation and which are independent.

Identifying such independent genes(features) whose expression patterns have meaningful biological links with phenotypic behaviours is important for knowledge discovery. In microarray analysis, biologists objective is to discover a small number of features which explains the behaviour of microarray data (Kumar and Valsala, 2013). Selective meaningful biomarkers from microarray data are important for patient stratification and for the development of personalized medicine strategies (Huang et al., 2015). From a machine learning point of view controlling the number of features helps to reduce over-fitting which lead to a better prediction of target variable on training data. The dimensionality of the feature space challenges about building a model and questions the effectiveness of knowledge discovery. Therefore, 10:1 per-class sample-to-feature ration is recommended for the creation of robust classifiers and predictive models (Kanal and Chandraskekaran, 1971).

The reason behind the feature selection is that classifiers trained on reduced feature space are more robust and reproducible than classifiers constructed on the original large feature space. In feature selection, we particularly search for features or correlated features. The features which do not provide useful information are called irrelevant features and the features which do not provide more information than the currently selected features are called redundant features (Kumar and Minz, 2014). The features which are not related or uncorrelated to class variables are called noise which actually introduces bias in prediction and reduce classification performance. Hence, noise should be handled for improving the performance of prediction and it can be made possible with dimensionality reduction. It can be achieved by either Feature extraction or by Feature selection (Drotár et al., 2015).

In feature extraction, new features are derived from the original input by choosing a new basis for the data. Feature selection helps in reducing the effect of high dimensionality on the dataset by finding the subset of features which will effectively define the data. Directly evaluating the subset of features becomes the NP-Hard problem (Chandrashekar and Sahin, 2014). To handle this issue we try to use a suboptimal procedure with tracable computations. We need to take care of another major issue where the feature is dependent on response variable rather than on predictors. Feature subset selection enables classifiers to focus on important features whilst ignoring the possible misleading features. From a computational complexity point of view, having a parsimonious set of features involved in the classification process helps in quickly scaling many learning algorithms with additional features (Dunne et al., 2002).

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

A better feature selection algorithm should always provide benefits such as insight into data, better classifier model, enhance generalization and identification of irrelevant features. It should also help in understanding the relation between features and target variables, reducing the computational requirement for solving a particular problem, efficient dimensionality reduction in case of high dimensional datasets where the number of observations are less than the number of features, it can help in improving the predictor performance which is used to solve a particular problem and increase the efficiency in terms of cost and time. Feature selection process contributes to knowledge discovery where discovered features can be used directly in future research. In bioinformatics, identification of important features can suggest new metabolic pathways and helps in identifying the hidden connection between specific cellular processes (Dunne et al., 2002).

Stability of a feature selection algorithm produces consistent feature subset, when new training samples are added or removed (Xin et al., 2015). A feature selection algorithm is stable only when it produces similar features under the training data variation. Ignoring the stability issue of the feature selection algorithm may draw a wrong conclusion. Among the highly correlated features, discarding the features which are correlated to the selected features but still associated with the response variable is one of the main cause of instability (Kamkar et al., 2015). A problem is said to be ill-poses if a small change in the input information causes a large change in the output (Cui et al., 2019). Instability with respect to the input data produces widely different output and makes the solution unreliable. The idea of Regularization converts an ill-posed problem in the stable form. Regularization modifies learning algorithm in such a way that it reduces generalization error but not training error.

The motivation of stability comes from increasing the confidence of domain expert in the analysis of result and select features which are relatively robust to the perturbation of input data (Kalousis et al., 2007). Stability provides the best objective criteria so that we can choose our feature selection algorithm, which will provide high-quality feature subset and also provide higher confidence in better classification performance. Strengthening of feature selection method with parallel analysis of stability develops high-quality feature subset (Goh and Wong, 2016). In knowledge discovery, stability plays an important role in feature selection to identify important features (George and Cyril Raj, 2015). A feature selection algorithm selects different subsets under perturbation of input data though most of these subsets are equivalent in terms of the classification performance (Li et al., 2015). Such unsteadiness reduces the assuredness of experts in the validation of selected features. Therefore it is very important to build a tough method to select reliable and significant features which are strong against the selection bias (Ambroise and McLachlan, 2002).

In the stability, the performance of the learning algorithm is used as an objective function because the stability of feature selection technique does not talk much about the confidence of the selected features. However, less stability does not imply low classification rate in every case (Somol and Novovicová, 2010). Stability helps in the trade-off between bias-variance of the classification error rate (Geman et al., 1992). Stability estimator of both the feature selection algorithm and classification algorithm does not create a boundary between instability of feature selection algorithm and classification algorithm. This issue can be tackled by the notion of preferential stability (Chen et al., 2019). Theoretically, a trade-off between bias-variance decomposition of feature selection error proposes that to get more stable features we do not have to sacrifice predictive accuracy. A better trade-off between bias-variance lead to more stable results with improved accuracy based on selected features. Margin-based instance weighting variance reduction is a better approach to achieve a better trade-off between bias-variance (Han and Yu, 2012). Margin-based instance weighting technique apply weight to each sample in a training set based on its influence to estimate the feature relevance. The hypothesis margin is used to measure the feature relevance at a given instance. Finally, the weighted training set is given as an input to the feature selection algorithm to select important features.

Following factors are responsible for the stability of feature selection algorithm: Dimensionality of the dataset (m), Number of the selected features (k), Sample size (n), Variance of the data, Symmetry of measurement where the stability value of the algorithm should be insensitive to the order of the result, Criterion used for feature selection and complexity of the feature selection algorithm (Loscalzo et al., 2009). Apart from these factors, there are other factors that cause instability, such as: Designing an algorithm without considering the stability, The existence of multiple sets of true markers and The Curse of dimensionality wherein few numbers of samples over thousands of numbers of features is a great source of instability (He and Yu, 2010).

2. Feature selection techniques

In this section, we have discussed the various feature selection techniques present in the literature.

2.1. Filter-based feature selection

The important characteristics of the data are used to assess the importance of feature for addition in the subset of features (Khoshgoftaar et al., 2013). This technique is alienated into two different categories: Rank Based and Subset Evaluation Based. Rank based category uses some univariate statistical techniques to evaluate the rank of each individual feature without considering the interrelationship between features (Yang and Mao, 2011). This technique flops to identify redundant features. Subset Evaluation Based category uses multivariate statistical techniques to evaluate the rank of the entire feature subset. The advantage of the multivariate statistical technique is, it consider feature dependency, no need of classifier and it is more efficient than wrapper technique in terms of computational complexity. The main drawback of the multivariate technique is, it slower and less stable as compared to the univariate ranking technique. Joint Mutual Information and Maximum of The Minimum Nonlinear Approach filter techniques produces the best trade-off between accuracy and stability (Bennasar et al., 2015).

2.2. Wrapper-based feature selection

This technique incorporates supervised learning algorithm in the process of feature selection. It ranks features based on the subset evaluation technique. Correlation and dependencies between the features are considered while selecting the features. Considering the bias of the prediction algorithm helps in optimizing the performance of the algorithm. In support vector machine (SVM), weight is assigned to each feature during the learning of SVM (Zheng et al., 2019). The main drawback of the wrapper technique is computational expensiveness due to searching of the optimal set from large space of dimensionality. Wrapper technique has a high risk of overfitting. SVM- Recursive Feature Elimination (RFE) and Greedy Forward Selection (GFS) strategy are some examples of the wrapper method.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

2.3. Embedded technique

The optimal feature subset is searched while building a classifier. Method of selecting optimal feature subset is specific for the given classification algorithm. Advantages of the embedded technique are the same as the wrapper technique but it is better in term of computational complexity than the wrapper technique. Lasso regression (Cynthia et al., 2019; Kang and Huo, 2019) and elastic net (Zou and Hastie, 2005; Xiao and Biggio, 2015) are some embedded techniques.

3. Feature searching strategies

In this section, we have discussed the different feature selection strategies which are used by different feature selection techniques present in the literature.

3.1. Forward sequential selection (FSS)

The objective of FSS is to create the optimal feature subset and ignore irrelevant and insignificant features (Wan, 2019). It searches for the best feature in every iteration and added to an empty set of optimal features. If all features are already added or if there is no improvement after adding any further feature, the search stops and returns the current optimal set of important features.

3.2. Backward sequential selection (BSS)

The objective of BSS is to consider the contribution of all features in the beginning and then tries to remove the most irrelevant and redundant features leaving a smaller optimal feature subset (Wan, 2019). It searches for the feature to be removed in every iteration from full dataset. The subsequent set is evaluated by some validation procedure. If the evaluation rate of new feature subset is better than the previous subset, then it replaces the current best feature subset. This process is continued until every feature is removed from the dataset and reaching an empty set. BSS outperforms the FSS in terms of computational performance.

3.3. Hill climbing (HC)

In HC either add or remove a feature from the dataset at a time. It searches the optimal features from the random set of features and then toggles the current status of each feature in the subset. The stopping criteria is set by defining the number of iteration for the selection of the optimal set. After reaching the limit of the last iteration returns the last optimal set of features (Wan, 2019).

4. Properties of feature selection stability measures

The stability estimator is used to calculate the robustness of feature selection algorithm to the input data perturbation by taking the average similarity of all pair of selected subsets. The main challenge of stability measure is whether the metric of stability measure make sense when feature selection algorithm produces feature subsets of different cardinalities. Every stability measure should satisfy the given properties (Nogueira and Brown, 2016):

4.1. Fully defined

Sometimes feature selection procedure produces different size of selected feature set when we iterate procedure n times. A good stability measure should always consider this property.

4.2. Upper and lower bounds

For a better understanding of stability measure, the value of stability measure should be in a finite range. Suppose defined range of stability measure is $[-\infty, +\infty]$, then output value 0.9 will be meaningless.

4.2.1. Deterministic selection → maximum stability

In Fig. 1 (left) (Nogueira and Brown, 2016) shows the different stability value of Lustgarten's measure for the different number of selected feature, where other methods showing the constant value of maximum stability for a different number of selected features.

4.2.2. Maximum stability → deterministic selection

In Fig. 1 (right) (Nogueira and Brown, 2016) Wald's measure and CWrel returns a constant value of maximum stability i.e. 1, for a different number of selected feature. Other methods show different stability values for a different number of selected features.

4.3. Correction for chance

This property ensures that when the feature selection procedure selects a random number of features their estimated stability value should be constant. Suppose, if the procedure P1 selects 5

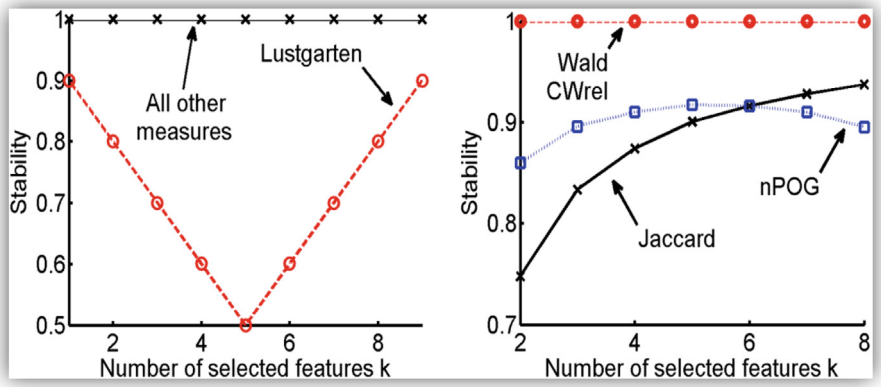

Fig. 1. Stability value of Lustgarten's measure for a different number of selected feature [LEFT]. Wald's measure and CWrel violate Property by giving constant stability over the different number of selected feature [RIGHT].

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

Table 1
Properties of Stability Measures.

Stability Measures	Fully Defines	Bounds	Maximum	Correction For Chance	Monotonicity
Jaccard	✓	✓	✓		✓
Hamming	✓	✓	✓		✓
Dice	✓	✓	✓		✓
POG	✓	✓	✓		
Kuncheva		✓	✓	✓	
nPOG	✓		✓	✓	
Wald	✓			✓	✓
CW_{rel}	✓	✓		✓	✓
Tanimoto	✓	✓		✓	✓
Symmetrical	✓	✓		✓	✓
Canberra	✓	✓		✓	
Spearman	✓	✓		✓	✓
Pearson	✓	✓		✓	✓

features and procedure P2 selects 6 features, the estimated stability value should be equal.

4.4. Monotonicity

Larger the intersection between feature subsets, greater the stability (Nogueira, 2018). Properties of different stability measures are given in Table 1.

5. Stability measures

The output of the feature selection algorithm can be in the form of a weighted score of each feature, the ranking of each feature or a subset of important features. These are called evaluation criteria of the feature selection algorithm. This evaluation criterion is divided into two parts (Mostafa et al., 2019):

- **Individual evaluation:** In this ranking of features is assigned according to its degree of relevance.

  A weighted scoring: $w = (w_1, w_2, \ldots\ldots, w_m)$, $w \in W \subset R^m$
  A ranking: $r = (r_1, r_2, \ldots..., r_m)$, $1 \leq r_i \leq m$

- **Subset evaluation:** In this feature subsets are constructed using the search strategy. Subset generation is the heuristic-search in which each state specify a feature subset for evaluation in search space.

Subset of features: $s = (s_1, s_2, \ldots\ldots, s_m)$, $s_i \in \{0,1\}$ where, 0 indicate absence and 1 indicate presence of feature.

Stability measures are divided into three categories (Mohana, 2016):

- Stability by Index/Subset (S_S)
- Stability by Rank (S_R)
- Stability by Weight (S_W)

5.1. Stability by Index/Subset (S_S)

The selected subset of features is represented as a binary vector of size 'm' where 0 represent absence and 1 represent the presence of the feature. The stability is calculated by the amount of overlap between the overall subset of selected features. Measurements of stability by index are given as:

5.1.1. Hamming distance (HD)

This calculates the amount of overlap between the two subsets (Mohana, 2016). It works with the binary vector of the selected feature subset. For larger m, $H(S_i, S_j)$ becomes smaller which leads to the more stable algorithm.

$$H(S_i, S_j) = \sum_{k=1}^{m} |S_{ik} - S_{jk}| \tag{1}$$

Hamming distance has the same impact on the stability result whether the feature S_i is selected in cross-validation or not. This will create less confidence about stability especially when the number of selected features are very few as compared to the overall dimensionality of the dataset. For total W feature subsets total hamming distance H_t is given as:

$$Ht = \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} H(S_i, S_j) \tag{2}$$

The total stability of all pairwise feature subset in S is defined by Average Normalize Hamming Distance (ANHD) (Mohana, 2016).

$$ANHD(S_i, S_j) = \frac{2 * H_t}{n * |W| * (|W| - 1)} \tag{3}$$

ANHD have results in the interval of (Mohammadi et al., 2016). 0 indicate the algorithm is most stable and 1 indicate the algorithm is not stable at all. The drawback of ANHD is it cannot deal with the different size of selected features. ANHD failed the property of correction for the chance, therefore it can deceive the result.

Normalize Hamming Index (NHI) is represent by:

$$NHI(S_i, S_j) = 1 - \frac{H(S_i, S_j)}{m} \tag{4}$$

The total stability of all pairwise feature subset in W is defined by Average Normalize Hamming Index (ANHI).

$$ANHI(S_i, S_j) = \frac{2 * \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} NHI(S_i, S_j)}{|W| * (|W| - 1)} \tag{5}$$

The value of ANHI represents the variation in the selected feature subsets. A higher value of ANHI gives more information about variation in feature subsets. Hamming distance measures failed in case of counting the intersection between two subsets.

5.1.2. Dice-Sorensen's index (DSI)

It calculates the overlap between two selected feature subsets (Mohana, 2016).

$$Dice(S_i, S_j) = \frac{2|S_i \cap S_j|}{|S_i| + |S_j|} \tag{6}$$

DSI give a result in the range of (Mohammadi et al., 2016). 0 indicates the two subsets are totally disjoint and 1 indicates the two subsets are identical to each other. DSI sometimes give slightly better and meaningful stability results because they are not divided by the union of subsets. On the other hand tanimoto distance and jaccard's index is divided by union of subsets.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

5.1.3. Tanimoto distance (TD)

Tanimoto distance calculates the amount of overlap between the two subsets of features and produces the value in the range of (Mohammadi et al., 2016). 0 indicates the two subsets are totally disjoint and 1 indicates the two subsets are identical to each other. It is the generalized version of Jaccard's index (Mohana, 2016).

$$T(S_i, S_j) = 1 - \frac{|S_i| + |S_j| - 2|S_i \cap S_j|}{|S_i| + |S_j| - |S_i \cap S_j|} \tag{7}$$

5.1.4. Jaccard's index (JI)

Jaccard's index (Mohana, 2016) measures the average similarity from all pairwise selected feature subsets (W).

$$J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{8}$$

$$JS = \frac{2}{|W| * (|W-1|)} \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} J(S_i, S_j) \tag{9}$$

The stability index (J_S) gives result in the rage of (Mohammadi et al., 2016) where value close to 0 indicates feature selection algorithm is unstable and value near 1 means algorithm is stable. The number of samples in the dataset influence the Jaccard's Index. Jaccard index can take correlation of features into account using:

$$JCi(K) = \frac{|S_i \cap S_j| + fC_i}{K} \tag{10}$$

K = Cardinality of S_i and S_j

fC_i = Sum of correlation values between dissimilar features.

For P selected subsets of data, stability is given by:

$$\bar{JC}(K) = \frac{\sum_{i=1}^{P} JC_i(K)}{P} \tag{11}$$

Both TD and JI give higher result when k = m. TD and JI are efficient as compared to DSI when the selected feature subsets have different cardinalities. They do not consider the dimensionality of the dataset (m) while calculating the similarity, but they compromise the number of selected features (k) in the measurements.

5.1.5. Kuncheva index (KI) or consistency index (I_C)

Because of the large size of selected feature subsets, stability estimators indicates higher overlap between the feature subsets. To overcome this drawback KI uses the correction term which discards the intersection by chance between the two selected feature subsets (Kuncheva, 2007). This also called a consistency index (I_c).

$$Ic(S_i, S_j) = \frac{|S_i \cap S_j| * m - k^2}{k * (m - k)} \tag{12}$$

The result of KI is in the range of [−1, 1]. 1 indicates the subset S_i and S_j are identical. −1 indicates two subsets have no intersection. 0 indicate for the independently drawn list. Average of all pairwise consistency indices is taken to calculate the consistency of more than two subsets.

$$AIc = \frac{2}{|W| * (|W-1|)} \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} I_c(S_i, S_j) \tag{13}$$

New similarity measure has been introduced to improve I_c:

$$Sa(S_i, S_j) = \frac{|S_i \cap S_j| - \frac{|S_i| * |S_j|}{m}}{\min(|S_i|, |S_j|) - \max(0, |S_i| + |S_j| - m)} \tag{14}$$

The result of S_a is in the range of [−1, 1]. 0 value indicates the stability of independently drawn random features, a positive value indicates the particular feature selection method is stable and negative value indicates the method is unstable. Adjusted stability measure (ASM) is the new stability measure which combines the result of multiple measures (Lustgarten et al., 2009). It can calculate the stability of subsets of unequal sizes.

$$ASM = \frac{2}{|W| * (|W-1|)} \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} S_a(S_i, S_j) \tag{15}$$

The intersystem similarity measure is very helpful in evaluating the similarity between the different feature selection method outcomes. It provides the diversity of feature selection methods. It compares the behaviour of two different feature selection methods on the same input dataset and sometimes compares two feature selection methods on two different datasets with the same feature selection setting.

5.1.6. Percentage of overlapping Gene/Features (POG)

It measures the consistency of selected feature subsets by counting the amount of intersection between selected feature subsets. POG is non-symmetric i.e. $POG(S_i, S_j) \neq POG(S_j, S_i)$ (Mohana, 2016). It will be symmetric if $|S_i| = |S_j|$

$$POG(Si, Sj) = \frac{|S_i \cap S_j|}{|S_i|} \tag{16}$$

POG matrix measures the consistency between the differentially expressed genes. The drawback of POG is it did not consider the correlation between features. To overcome the drawback of POG, POGR has been introduced (Mohana, 2016).

$$POGR(S_i, S_j) = \frac{|S_i \cap S_j| + Z}{|S_i|} \tag{17}$$

Z = Number of genes in S_i that are absent in S_j but considerably positively correlated to at least one gene in S_j.

Z captures the correlation between features and consider such features as a single feature. Normalize version of POG and POGR discards the dependency between the results.

$$nPOG(S_i, S_j) = \frac{|S_i \cap S_j| - E(|S_i \cap S_j|)}{|S_i| - E(|S_i \cap S_j|)} \tag{18}$$

$$nPOGR(S_i, S_j) = \frac{|S_i \cap S_j| + Z - E(|S_i \cap S_j|) + E(Z)}{|S_i| - E(|S_i \cap S_j|) - E(Z)} \tag{19}$$

$E(|S_i \cap S_j|)$ = Expected value of the shared feature

E(Z) = Number of features in the S_i which are not shared but positively correlated with features in S_j.

POG and POGR are bunded by the interval (Mohammadi et al., 2016). Similarly, nPOG and nPOGR are bounded in the interval of [−1, 1]

5.1.7. Consistency measures (CM)

Stability value produces by different stability estimators on the same system are bounded in different ranges, this makes them hard to compare (Somol and Novovicová, 2010). Most of the available measures are applicable only for the feature selection problem with the prespecified size of a subset (k). To overcome the above issue new modified stability measure has been introduced. The designed stability measure produce result in the range of (Mohammadi et al., 2016). Stability measure of value 1 represents the stable feature selection algorithm and value 0 represents a low level of feature selection algorithm stability. This evaluation of stability is based on the frequency of feature occurrences. X is the subset of Y representing all features in S. F_f is the occurrence of feature f in system S. N is the number of occurrence of any feature in system S. The minimum value (F_{min}) of occurrence of feature f is 1 and the maximum value (F_{max}) is 'm'. Stability value of the feature $f \in X$

is bounded in the range of (Mohammadi et al., 2016). 0 indicates feature f is present in only one subset among the n subsets in system S. 1 indicates feature f exists in every subset of the system S. Consistency (C(f)) of the feature f in the system S is given as:

$$C(f) = \frac{F_f - F_{min}}{F_{max} - F_{min}} \tag{20}$$

To define the consistency of the whole system

$$C(S) = \frac{1}{|X|} \sum_{f \epsilon X} C(f) \tag{21}$$

This measure overemphasizing the lower frequency features, therefore weighted consistency CW(S) has been introduced (Lustgarten et al., 2009).

$$CW(S) = \sum_{f \epsilon X} \frac{F_f}{N} * \frac{F_f - F_{min}}{F_{max} - F_{min}} \tag{22}$$

Whenever m > |X|, it indicates that features are present in more than single subset and therefore CW(S) > 0. The selected feature subsets automatically get more similar to each other when the size of the selected feature subsets comes closer to the actual size of the dataset. In this situation applying CW(S) for various feature selection methods may yield the less confident result. Producing the system of differently sized subsets is called the problem of subset-size bias. Relative weighted consistency (CW_{rel}) tackle this problem by suppressing the effect of the size of subsets in a system.

$$CW_{rel}(S, Y) = \frac{CW(S) - CW_{min}(N, n.Y)}{CW_{max}(N, n) - CW_{min}(N, n.Y)} \tag{23}$$

$$CW_{min}(N, n, Y) = \frac{N^2 - |Y|(N - D) - D^2}{|Y| * N(n - 1)} \tag{24}$$

$$CW_{max}(N, n) = \frac{H^2 - (N - D) - H * n}{N * (n - 1)} \tag{25}$$

D = N mod |Y|
H = N mod m
CW_{rel} incorporates randomness into the feature selection.

5.1.8. Symmetrical uncertainty (SU)

SU is an entropy-based nonlinear correlation (Mohana, 2016). It takes feature value in the account while calculating stability, not the feature indices. SU identifies the correlated features in all selected subsets. Information gain, IG (S_i | S_j) = IG (S_j|S_i) this property makes SU a symmetric measure. SU has one undesirable property of not bounding by any constant.

$$SU(S_i, S_j) = 2\left[\frac{IG(S_i|S_j)}{H(S_i) + H(S_j)}\right] \tag{26}$$

IG = Information Gain = $H(S_i) - H(S_i|S_j)$
$H(S_i)$ = Entropy = $\sum_{x \in S_i} p(x) * \log_2(p(x))$
$H(S_i|S_j) = \sum_{y \in S_j} p(y) \sum_{x \in S_i} p(x|y) * \log_2(p(x|y))$

The computation of IG for every pair of selected features make SU computationally expensive. The result of SU influence by the number of selected features (k) and the awful result arises when k = m.

5.2. Stability by rank (S_R)

The correlation between features is evaluated to quantify the stability of feature selection method using feature ranking. The main drawback of these measures is they cannot handle subsets of features with different cardinality.

5.2.1. Spearman's rank correlation coefficient (SRCC)

Stability of two ranked sets of features R_i and R_j is given by:

$$SRCC(R_i, R_j) = 1 - 6 \sum_{t=1} \frac{(R_{it} - R_{jt})^2}{m(m^2 - 1)} \tag{27}$$

The value of SRCC is in the range of [−1, 1]. Two ranks of features are identical when SRCC value is 1 and exactly inverse order when −1. 0 indicate no correlation between R_i and R_j. The overall stability of all feature subsets is:

$$ASRCC(R_i, R_j) = \frac{2}{|W| * (|W - 1|)} \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} SRCC(R_i, R_j) \tag{28}$$

5.2.2. Canberra distance (CD)

This measure shows the absolute difference between two rank sets (Mohana, 2016). Value of the CD is directly proportional to the number of features. Higher the value of m, larger the value of the CD.

$$CD(R_i, R_j) = \sum_{t=1}^{m} \frac{|R_{it} - R_{jt}|}{R_{it} + R_{jt}} \tag{29}$$

Weighted version of CD can be define as:

$$WCD^{(k+1)}(R_i, R_j) = \frac{\sum_{t=1}^{m} |\min\{R_{it}, k+1\} - \min\{R_{jt}, k+1\}|}{\min\{R_{it}, k+1\} - \min\{R_{jt}, k+1\}} \tag{30}$$

The result of CD is bounded in between (Mohammadi et al., 2016). Top k features are considered as the most important features. Both CD and WCD is normalized when divide by m.

5.3. Stability by weight (SW)

These measures consider the weight of feature set f while calculating the robustness of feature selection algorithm. It takes the two sets of weight W_i and W_j for the complete feature set in datasets and returns the correlation between them as the stability. The main drawback of these measures is it cannot deal with the subsets of features of different sizes.

5.3.1. Pearson's correlation coefficient (PCC)

PCC calculates the correlation between the weights of the selected subsets of the features (Geman et al., 1992). PCC returns the result in a range of [−1, 1]. 1 means weight vector perfectly correlated and −1 indicate weight vectors are anti-correlated. 0 indicates no correlation between weight vectors. For a larger number of features weight approach to 0 indicates higher stability. PCC is the symmetric stability measure.

$$PCC(W_i, W_j) = \frac{\sum (W_{it} - \mu_{W_i})(W_{jt} - \mu_{W_j})}{\sqrt{\sum (W_{it} - \mu_{W_i})^2 \sum (W_{jt} - \mu_{W_j})^2}} \tag{31}$$

μ = Mean of the feature set f

In (Geman et al., 1992) different feature selection algorithm applied on a dataset for getting ranks, weights and subsets of features. Stability evaluation by using weights of feature (S_w) provides a better understanding as compared to using the ranking of features (S_R) because it uses actual feature coefficients. Highest stability is given by S_w. Stability value using subsets (S_S) does not correlate with the other two measures. High cardinality of the selected feature subsets indicates the more probability of features in common. Therefore stability value will also increase.

A novel stability estimator is proposed in Zheng et al. (2019) which satisfy all the desired properties of stability measure. This novel stability measure is given as:

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

$$\Phi(z) = 1 - \frac{\frac{1}{m}\sum_{i=1}^{m} f_i^2}{\frac{\bar{k}}{m}\left(1 - \frac{\bar{k}}{m}\right)} \tag{32}$$

$$f_i^2 = \frac{N}{N-1} P_f(1 - P_f) \tag{33}$$

$\bar{k}$ = Average number of features selected over the N feature set in z.

f_i^2 = Sample variance of z_f.

The value of estimator is in the range of (Mohammadi et al., 2016).

6. Solution to instability

Till date, we have various methods to solve instability of feature selection algorithm. In this segment, we tried to cover all methods present in the existing literature. Fig. 2. Summarise the different methods to solve the different source of instability.

6.1. Feature information strategy

Feature information strategy measures the significance of each feature based on some assessment standards like the accurate measure of class variable. Then stable features have been selected from these highly important features (Liu et al., 2017). Feature Importance in Nonlinear Embedding (FINE) approach is used for the ranking of features based on their contribution to accurate classification in low dimensional space. This low dimensional feature space is achieved via Non-Linear Dimensionality Reduction (NLDR) (Ginsburg et al., 2016). Since features in low-dimension are less sensitive to small data perturbation, feature ranking is more stable than traditional filter method.

Multi-knockoff procedure guarantees False Discovery Rate (FDR) control and has better statistical properties than single knockoff procedure even when the number of prominent features is small (Gimenez and Zou, 2018). Knockoff procedure allows us to discover important features while controlling the FDR. The advantage of the knockoff is that if we have a good model of feature X then we can identify significant features without considering how output Y depend on feature X. By averaging over the K multi-knockoff decrease the threshold of the minimum number of rejection which leads to improvement in power and stability.

Extracting hybrid-features is better than extracting basic-features because it contains in-flow behavior of features. In-flow behavior characteristics of traffic flow have been analyzed to select important features from traffic data of mobile app (Liu et al., 2019). A metric is designed to measure the degree of drift of flow features. Based on this a composite metric ranks a feature. Discrimination power and Degree of drift evaluation based Feature Selection (DDFS) algorithm is designed based on the above metrics to discover the discriminative and stable features. DDFS selects features of high discrimination capability but a lower degree of drift.

SVM-REF is a multivariate iterative backward feature selection method. It takes interaction between features into account while assessing the relevance of the features (Lahmiri and Shmuel, 2019). To improve the stability against the input data perturbation, 9-fold cross-validation is used. The Cumulative Ranking Score (CRS) of all features is calculated in every iteration, which is used to compute the importance of each feature in creating a difference between classes. This parameter combined the ranking of the features gained from different subsets. The features having high cumulative ranking are robust and exact set of genes which are responsible for the disease.

6.2. Feature relatedness

Feature relatedness computes the connection between the features. Covariance-lasso (C-LASSO) calculate the resemblances between features with the help of feature covariance matrix. It addresses the instability of L_1-norm. The objective function of C-LASSO is given as (Kamkar, 2016):

$$J_{\beta,\Omega}^{\text{argmin}}(\beta,\Omega)\frac{1}{2} = ||y - X\beta||_2^2 + \lambda||\beta||_1 + \frac{\eta}{2}\beta^T\Omega^{-1}\beta \tag{34}$$

Such that, $\Omega \geq 0$, $\text{tr}(\Omega) = 1$

Ω = Covariance matrix

λ, η = Tunning parameters

Model fitting and sparsity are the essential components for the accuracy of the regression model. Tunning parameters balancing the trade-off between these two components. Cohen's kappa coefficient is used to measure the similarity between the two feature sets. Covariance SVM (C-SVM) identifies the correlation between the features using convex objective function and select relevant features (Kamkar et al., 2015). Combination of SVM with Elastic net penalty yields new regularization formation to find a connection between features based on their relatedness (Ye et al., 2011).

Max-Min Correntropy Criterion (MMCC) is a new formulation explored for the feature selection. Correntropy is a local similarity measure in Information Theoretic Learning (ITL)

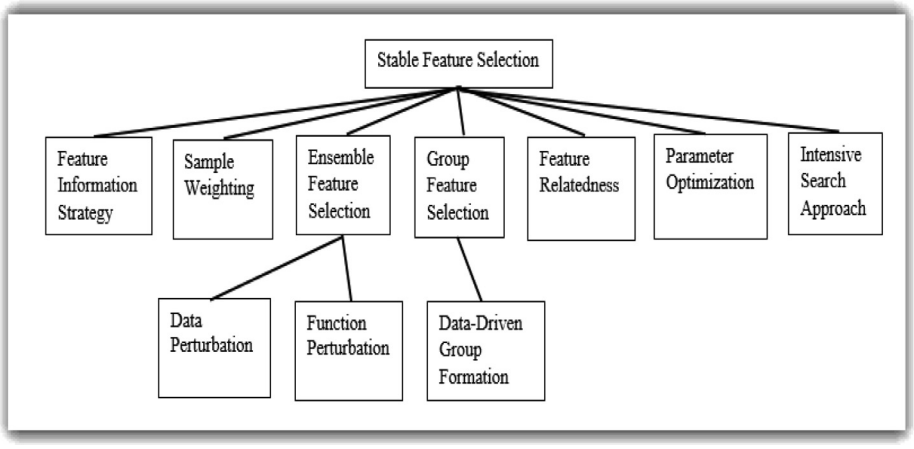

Fig. 2. A network of stable feature selection methods.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

(Randall et al., 2019). Kernel width influences the performance of Correntropy, smaller kernel width is responsible for data loss. In contrary, high kernel width makes Correntropy weak against the high corruption and non-gaussian noise. The ANHD value of MMCC almost reached to 0, which indicates the robust nature of MMCC against the noise and outliers (Storn and Price, 1997). MMCC algorithm does not have 'out of memory' issue (Brest et al., 2006).

The combined network of synergistic proteomics is used to overcome its passive coverage and consistency issue (Bensimon et al., 2012; Goh et al., 2012). Ranked-Based Network Algorithms (RBNAs) is a network-based algorithm which proved its utility in case of selecting important features with high stability (Selvaraj et al., 2018). Three benchmark approaches of RBNAs has been introduced: 1. SNET (SubNet) (Liu et al., 2019) 2. FSNET (Fuzzy SNET) (Soh et al., 2011) and 3. PFSNET (Paired FSNET) (Lim and Wong, 2014). In SNET, if the protein g_i is among the top n% most influential protein in the tissue p_k then $f_s(g_i, p_k) = 1$, otherwise $f_s(g_i, p_k) = 0$. FSNET is similar to SNET except for function $f_s(g_i, p_k)$ is assigned a value between 1 and 0. In PFSNET, $f_s(g_i, p_k)$ define score $\delta(S, p_k, X, Y)$ for complex S and tissue p_k with respect to the classes X and Y.

$$\delta(S,\ p_k,\ X,\ Y) = \text{Score}(S,\ p_k,\ X) - \text{Score}(S,\ p_k,\ Y) \tag{35}$$

Subgroup based Multiple Kernel Learning (MKL) is used for classification of biomedical image texture datasets. The main objective of MKL is to observe its performance incorporate stability in feature selection (Fernandez-Lozano et al., 2015). MKL eliminate features with similar characteristics and selects only features which increase the understanding of the results. MKL perform for multiple feature selection. Grouping of features by an external criterion allows to include a single feature which is represented by its own base kernel. It also allows the MKL algorithm to select the correct kernel parameter. This will select the most representative features of the problem.

Unsupervised Graph Self-Representation Sparse Feature Selection (GSR-SFS) combined subspace learning with feature selection method which improves the interpretability of features (Zhu et al., 2013; Hua et al., 2017). In this method, feature level self-representation loss function and $L_{2,\ 1}$-norm regularization term, represent every feature by its relevant features. The idea behind the feature level self-representation loss function is that the most powerful feature has more chances to represent other features jointly. In contrary, the less powerful feature has no chance to represent other features. The $L_{2,\ 1}$-norm regularization term penalizes all coefficient in order to joint selection or rejection of features for the prediction of the response variable. Graph regularisation term conducts subspace learning and improves stability by maintaining the original structure of data into low-dimensional space. In self-representation graph low-rank dimensionality reduction (SGLR), low-rank constraint and a graph Laplacian regularizer are used to conduct subspace learning in unsupervised feature selection (He et al., 2016).

LASSO is inconsistent when performed on correlated data because it allocates a nonzero weight to only a single feature among a group of correlated features (Yuan and Lin, 2006). Flexible factor modelling for the covariance can create robust feature selector (Grollemund et al., 2019; Ramondta and Ramírezb, 2019). A supervised factor analysis model takes advantage of flexible parametrization and shows the dependency by existing latent factors. Dependence is capture in low dimensional space. After adjustment or latent effect, weak correlation applied to decorrelated data. Decorrelation in factor adjustment leads to better performance. LASSO select less number of features in the presence of factor adjusted data. Factor adjustment helps in blocking the effect of heterogeneity, improves stability and prediction error of selected variables.

6.3. Sample weighting

The principle behind this technique is to assign different weights to each sample based on the influence of sample on the feature relevance. The local profile of feature relevance is used to measure the influence of every sample (Li et al., 2017). Then the feature selection algorithm train on the weighted training set.

Feature weighting as Regularised Energy-Based Learning (FREL) is a new feature selection algorithm which is guided by Energy-Based Learning (LeCun et al., 2006). FREL predicted the stability under L_1 and L_2 regularization. Because of the nature of L_1-norm, the sparse solution is produced by the feature selection algorithm. Sparsity and stability are the different sides of the same coin for classification and regression problems (Xu et al., 2012). The sparse algorithm is not stable and they can identify redundant features. Therefore the sparse algorithm has a non-unique solution and thus may be ill-posed. Value of energy function measure the goodness of fit and decides the degree of compatibility of a model between input variable *x* and output variable y. A small value means highly compatible configuration and large value means highly incompatible configuration. The difference between the energy of the correct answer and the incorrect answer for x_i is defined as the generalize margin loss of the sample. FREL compute the feature weight and then convert them into feature rank. The feature with the large weight has a high rank and feature with less weight have low rank.

Maximum Relevance Minimum Redundancy (MRMR) have two different criteria: Mutual Information Difference (MID) and Mutual Information Quotient (MIQ). Balancing the trade-off between maximum relevance and minimum redundancy is necessary for the stable feature selection. A feature may have different weights for the relevance and redundancy in the feature selection. This weighting parameter helps in controlling the stability of MRMR (Gulgezen et al., 2009).

6.4. Parameter optimization

To solve the instability problem, the feature selection method is used in the course of the parameter optimization process. The idea is to select an active set of parameters which optimize the current optimization process. The author investigated the nonlinear regression model with the squared error function and the logistic regression model with the cross-entropy error function (Isachenko and Strijov, 2018). Newton method is used for nonlinear regression and Gauss-Newton method for model linearization. The Newton method for logistic regression brings Iteratively Reweighted Least Square (IRLS) algorithm (Isachenko and Strijov, 2018). The Quadratic Programming Feature Selection (QPFS) (Katrutsa and Strijov, 2017) is used to select an optimal set of parameters. QPFS identify the most impactful parameters on model residuals. The proposed algorithm achieves less error and more stability as compared to other methods.

6.5. Intensive search approach

Intensive search approaches include parallel search strategies and a Genetic Algorithm (GA) (Sakae et al., 2018). Most of the search processes stuck at potentially different local maxima, therefore, different search techniques improves the stability by increasing the scope of the search by considering most candidate mask at each decision point. In GA, the initial element is the binary vector of length N containing either 0 or 1. A strong set of features is produced by doing the reproduction of the highest performing variables. This process continues until we get the best set of features for the search problem (Sakae et al., 2018). Choice of parameters in the GA creates a huge impact on the performance. The significant size of

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

the population and a reasonable number of generations incorporate heavy runtime computational expenses in GA.

An Improved version of Grey Wolf Optimization (IGWO) technique has been used to select optimal features that determine the protein structures (Sharma and Gupta, 2018). Grey wolf optimization (GWO) is an evolutionary algorithm which mimics the hunting behaviour and leadership hierarchy of grey wolves (Mirjalili et al., 2014). The hunting behaviour of grey wolf includes hunting, looking, circumscribing the prey and attacking (Sharma and Gupta, 2018). The leadership hierarchy of the grey wolf is divided into 4 types- α, β, δ and Ω. The proposed IGWO technique gives maximum accuracy with ANN classifier.

6.6. Group feature selection

The idea behind this technique is to group the highly correlated features present in high-dimensional datasets which are resistance to the variations of training samples. The stability of the selection process can be improved, if we consider this group as a single entity (Li et al., 2017). Fig. 3 shows the process flow of group feature selection.

There are two key features in group feature selection: Feature Group Generation and Feature Group Transformation. Feature group generation identifies the groups of associated features. This can be done by Knowledge-driven methods or Data-driven methods. The knowledge-driven method requires deep domain knowledge to form groups and data-driven methods use information contained in input data for group formation. Feature group transformation produces a lucid picture of the feature group.

6.6.1. Data-Driven group generation

Data-driven group generation recognizes a group of features using either cluster analysis or density estimation (Jeitziner et al., 2019). Instead of relying on domain knowledge of biology they form groups based on information contained in input data. Group-lasso is applicable when different groups form by correlated features (Jacob et al., 2009). However, group-lasso found unsuitable when the features are naturally present in its tree structures. For this kind of features, tree-lasso is suitable for interpretability of features (Kamkar et al., 2015). Tree-lasso achieves correlated features in the form of hierarchical structure.

Tree – guided Recursive Cluster Selection (T-ReCS) method effectively selects group of features (Villaruz et al., 2015). T-ReCS improves predictive stability without compromising with accuracy. It cures the instability by selecting important features at the cluster level. T-ReCS can efficiently handle features which do not belong to a cluster (called orphan features) without having prior knowledge of cluster. Initial cluster formed with the help of Max-Min Parent Children (MMPC) algorithm (Lagani and Athineou, 2017). Size of the cluster is based on the threshold value defined by the user. A hierarchical tree structure pointing out towards the resemblance between variables. Leaf of the constructed tree represents a single variable and internal node represents a cluster of variables. A deeper node in a tree has a more similar pattern among its members. The tree structure is generated by Recursive K-means Spectral clustering (ReKS) which formed a tree with effective speed and gives a more balanced tree when applied to heterogeneous data (Huang et al., 2013).

Fused-lasso is used when features are highly correlated due to the ordering between them (Tibshirani and Saunders, 2005). Fused-lasso selects neighboring features and improves feature stability. For stable feature selection in the diagnosis of Alzheimer's disease a non-negative fused-lasso incorporates two important factors: Spatial cohesion of lesion voxels and a positive correlation between explanatory and response variables (Tichý et al., 2019). The model does not select negatively correlated features because of the additional non-negative constraints. To solve constraint optimization and prove its convergence, Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) is used. This procedure allows us to discover features which are truly responsible for Alzheimer's disease while controlling the false discovery rate.

A stable feature selection should provide consistency among repeated Electroencephalography (EEG) measurement of the same condition on the same subject over the course of time (Lan, 2018). The Intra-class Correlation Coefficient (ICC) is a stability estimator which describes the resemblance of data in the same group (McGraw and Wong, 1996). Larger ICC value indicates higher similarity among the group data. Features with higher ICC value are more stable and can better discriminate different emotions.

6.7. Ensemble feature selection

Ensemble learning technique works on the idea of "Wisdom of crowds" which contains that large groups of people are collectively smarter than even individual experts when it comes to problem-solving, decision making and predicting. Ensemble learning is a part of machine learning which effectively produces a robust and accurate learning solution (Wang and Chiang, 2011). Ensemble learning technique use bagging technique which takes the average of several learning techniques builds from random subsamples of the original dataset (Diren et al., 2019). Different partitioning of the training data selects different routes in the output. This problem can be solved by aggregating several runs of a sequential search (Sánchez et al., 2019). An ensemble solution to instability helps in stabilizing the process. It incorporates stability consideration into designing stage of the algorithm (Dessì and Pes, 2015). Fig. 4. Contain the outline of ensemble feature selection.

Diverse local learners can be constructed by Data perturbation and Function perturbation. Data perturbation uses bootstrapping, over-sampling, under-sampling to generate random subsamples

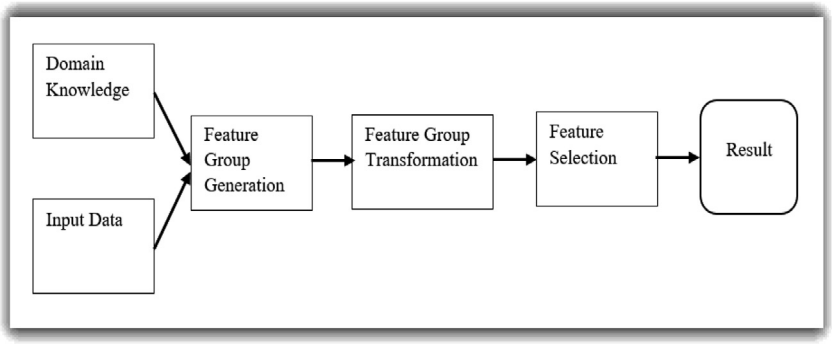

Fig. 3. Group feature selection framework.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

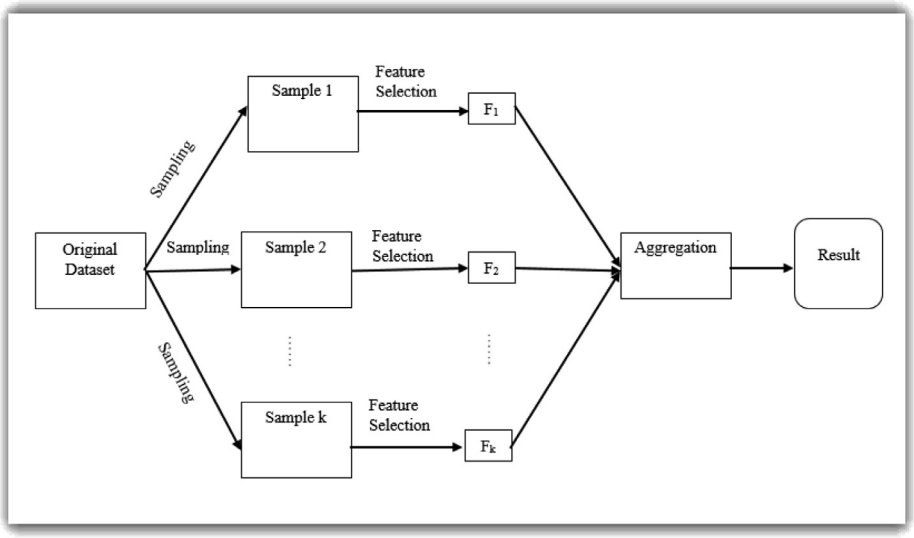

Fig. 4. Ensemble feature selection.

of the dataset and apply the feature selection technique on each generated subsample. In the end, an ensemble method combines the results generated by the same feature selection method (Chatterjee, 2019). In function perturbation, different feature selection methods are applied to the same datasets. Like in the ensemble method different feature ranking list is combined into a single set to get stable features.

6.7.1. Data perturbation

Survival Count on Random Subsamples (SCoRS) method helps in removing irrelevant features in functional Magnetic Resonance Images (fMRI) (Rondina and Hahn, 2014). This study is based on the survival frequency of features after several iterations instead of relying on the coefficient values given by the L_1-norm regression (Meinshausen and Buhlmann, 2010). Stable features are selected by the repetitive application of L_1-norm under data perturbation. The features that survive in a large fraction of perturbation even after several iterations are considered as important features. A threshold value is used to select topmost stable features. More significant features can be selected by the recombination of features in different subsets.

Recursive feature elimination (RFE) machine learning approach classify the samples by iteratively eliminating the least influential features. Performance benchmark is done by Precision, Recall and F-Score (Patil and Rao, 2018). Cross-validation is generally used for error estimation but it shows large variance when applied to small-sized data, therefore, it creates ambiguity in result (Braga-Neto and Dougherty, 2004). Multi-Criteria Fusion Based Recursive Feature Elimination (MCF-RFE) improves the stability by selecting the significant features which are less sensitive to the incorrect estimation of statistical parameters like mean, variance, standard deviation, etc (Du et al., 2016). MCF-RFE use score-based and ranking-based fusion methods to generate feature ranking (van Erp and Schomaker, 2000; Yan and Zhang, 2012).

Stability deficiency and balanced decision tree decrease the robustness of random forest. Selecting important features using an iterative procedure help to overcome this limitation (Park and Kim, 2015). Regularised Random Forest (RRF) and lasso are used for the classification of High-Dimensional Shape Description (HDSD) of brain morphometry (Wade et al., 2017). Lasso give more classification accuracy, however, RRF and No Feature Selection (NFS) gives a more robust performance.

Sequential Random K-Nearest Neighbor (SRKNN) is a wrapper technique based on the nearest neighbour ensemble classifier (Park and Kim, 2015). A non-hierarchical structure of the nearest neighbour classifier remove the instability and high variance occurred in the random forest approach. SRKNN select features based on the firmness of features and not on training accuracy. This property improves the stability of SRKNN in feature selection.

L_1-norm SVM efficiently removes irrelevant and insignificant features using backward feature elimination method based on feature ranking (Moon and Nakai, 2016). Ensemble selection for L_1-norm shows the remarkable score for stability as compared to other methods. To generate more converge results, bootstrap samples are drawn from training datasets using bagging technique. The regularization parameter of L_1-norm SVM is optimized for every generated bootstrap sample. The linear kernel is used as a kernel function because it is less prone to overfitting. A feature having coefficient value 0 is eliminated from the bootstrap sample and then cross-validation score has been recorded. This procedure is iterated until no feature has a coefficient value 0. Finally, a set of significant features is produced by adding all remaining features in the bootstrap sample.

The aggregation approach produces impressive stability improvement as compared to standard wrapper-based feature selection techniques because both FSS and BSS are deterministic and produces a similar output in every iteration on the same training data. After finishing all trials, the matrix of the collected mask is used to form the Aggregated Frequency Histogram (AFH). This histogram selects the most frequent features. A feature is selected from AFH based on some threshold value and those exceeded the threshold value are added in a final set of features.

Evolutionary Algorithms Feature Selection Stability Improvement System (EAFSSIS) is made up of two components: 1. Filter ensemble ranking 2. Feature selection method. Training data generated by cross-validation is given as an input to the filter ensemble ranking which generates several groups of samples. Then filter methods are applied on every generated sample to rank features. The final rank of the feature is yielded by the ensemble method. Finally, the ranking result along with the training data are given as input to feature selection method to select significant features.

6.7.2. Function perturbation

In function perturbation, averaging the outcome of different feature selectors on the same input dataset give the final result.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

Different feature selection techniques, instance-level perturbation, feature-level perturbation, stochasticity in the feature selector, etc. incorporate the variation in the feature selectors. Weighted voting can be used for the aggregation of different selection techniques.

A base feature selectors derived from FREL is trained on the bootstrap samples derived from the original training data (Tran et al., 2019). Averaging the outcome of base feature selectors gives the final rank of the feature. There are various functions of the rank aggregations (Haury et al., 2011; Abeel et al., 2010). Finally, each aggregation strategy gives the list of top features with the largest score and higher stability.

7. Discussions

We summarized the different modes of feature selection techniques and feature selection strategies in Section 1. Features selected by the wrapper-based and embedded techniques sometimes do not perform well with the other classifiers because they both utilize their own learning algorithm for feature selection. Computational complexity of filter-based technique is less as compared to embedded and wrapper-based techniques. The complex nature of wrapper-based techniques creates the high risk of overfitting. Filter-based methods provide more stable sets of selected features due to their robust nature against overfitting (Hastie et al., 2001). They generally use univariate or multivariate statistical techniques and independent of any learning techniques. Backward sequential selection feature selection strategy outperforms forward sequential strategy and hill climbing strategy in term of computational complexity. Hill climbing strategy is more optimal but its random nature increases its complexity.

Designing an algorithm without considering the stability, existence of multiple equally predictive feature subsets and the curse of dimensionality are the common sources of instability. The curse of dimensionality hinders the selection of stable features therefore research progresses in the related field will develop better stable feature selection algorithm. Other than above mentioned sources of instability, some properties of feature selection techniques such as number of selected variables, sample size, criteria used for feature selection and complexity also affects the stability.

In this review, we have discussed almost all stability measures based on index, rank and weight of the selected features. There are numerous techniques exists which is used to measure the stability based on index and rank but there are only few techniques to measure stability based on weights. Among the various stability measures, only Pearson's correlation coefficient computes the stability by considering the weight of the feature. Our main concern is that stability measures are unstable which means if we run different stability measure on the same feature selection algorithm, we will get different values of stability measures. Another problem with stability measures is that not a single measure fulfills all the required properties of stability measure.

By the extensive survey of finding a solution to the instability of feature selection algorithm, group feature selection and ensemble feature selection are the most widely used methods. Group feature selection was used because of the presence of highly correlated features in the high-dimensional dataset. However, a grouping of features reduces the instability by a small amount because of the reproducibility issue in the transformed phase. Ensemble feature selection provides a more general-purpose solution. Along with the function perturbation and data perturbation, hybrid perturbation in the ensemble feature selection applies the data perturbation in the initial stage and function perturbation in the final stage to improve the probability of getting stable feature selection. Using an ensemble strategy is not always beneficial but it can give better results even if selection technique is less stable.

8. Conclusions

Recent advancement in high-throughput technologies, radiomics, neuroimaging, sensors, etc. produces very high-dimensional data. With the proliferation of high-dimensional datasets in recent years, feature selection has received attention of researchers and data-mining professionals in terms of both performance and computational efficiencies. A feature subset selected by a feature selection technique is evaluated for relevance towards a task such as classification or knowledge discovery from high-dimensional feature space. One important characteristic of feature selection technique is stability. Stability is the insensitivity of a feature selection algorithm to a small perturbation in input training data. Our focus in this review paper is to address the problem of stability, its importance, various stability measures used to evaluate subsets of selected features and solutions to the different source of instabilities. Subsets of selected features may be generated using filter-based, wrapper-based or embedded techniques. These feature selection techniques use different feature selection strategies such as forward sequential search, backward sequential search and hill climbing to select most important features. In order to measure the stability of a feature selection techniques, a similarity measure is needed to assess the overlap of a pair of feature subsets. The strength and weakness of all similarity measures are presented. We summarized the solutions to the different source of instabilities based on feature information Strategy, feature relatedness, sample weighting, parameter optimization, intensive search approach, group feature selection and ensemble feature selection. From the extensive survey we can say that group feature selection and ensemble feature selection are the most widely used methods although using an ensemble strategy is not always beneficial but it can give better result even if selection technique is less stable. Stable feature selection is very important from theoretical as well practical perspective. More progressive research need to be developed to explore this challenging topic. Throughout this study, we discussed the current researches going on in feature selection techniques stability analysis within the domain of bioinformatics, image analysis, healthcare, business analytics, networking, etc and have identified the shortcoming of these works to explore possible opportunities for future work.

Declaration of Competing Interest

None.

References

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y., 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26 (3), 392–398. https://doi.org/10.1093/bioinformatics/btp630.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 99 (10), 6562–6566. https://doi.org/10.1073/pnas.102102699.

Bennasar, M., Hicks, Y., Setchi, R., 2015. Feature selection using Joint Mutual Information Maximization. Expert Syst Appl. 42 (22), 8520–8532. https://doi.org/10.1016/j.eswa.2015.07.007.

Bensimon, A., Heck, A.J.R., Aebersold, R., 2012. Mass spectrometry-based proteomics and network biology. Annu. Rev. Biochem. 81. https://doi.org/10.1146/annurev-biochem-072909-100424, pp. 18.1–18.27.

Braga-Neto, U.M., Dougherty, E.R., 2004. Is cross-validation valid for small-sample microarray classification? Bioinformatics 20 (3), 374–380. https://doi.org/10.1093/bioinformatics/btg419.

Brest, J., Greiner, S., Bošković, B., Mernik, M., Zumer, V., 2006. Self-adapting control parameters in differential evolution: a comparative study on numerical benchmark problems. IEEE Trans. Evolut. Comput. 10 (6). https://doi.org/10.1109/TEVC.2006.872133.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electrical. Eng. 40, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chatterjee, S., 2019. The scale enhanced wild bootstrap method for evaluating climate models using wavelets. Stat. Probab. Lett. 144, 69–73. https://doi.org/10.1016/j.spl.2018.07.020.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

Chen, G., Cao, M., Yu, J., Guo, X., 2019;Jan(461):. Prediction and functional analysis of prokaryote lysine acetylation site by incorporating six types of features into Chou's general PseAAC. J. Theor. Biol. 461, 92–101. https://doi.org/10.1016/j.jtbi.2018.10.047.

Cui, J., Peng, G., Lu, Q., Huang, Z., 2019. Modified special HSS method for discrete ill-posed problems and image restoration. Int. J. Comput. Math., 1–20 https://doi.org/10.1080/00207160.2019.1585827.

Cynthia, K., Lyne, L., Bareil, Céline, B., et al., 2019. Lasso regression for the prediction of intermediate outcomes related to cardiovascular disease prevention using the TRANSIT quality indicators. Med. Care. 57 (1), 63–72. https://doi.org/10.1097/MLR.0000000000001014.

Dessì, N., Pes, B., 2015. Stability in biomarker discovery: does ensemble feature selection really help? Int. Conf. Industr. Eng. Other Appl. Appl. Intell. Syst., 191–200 https://doi.org/10.1007/978-3-319-19066-2_19.

Diren, D.D., Boranlhsan, S., Selvi, H., Hatipoglu, T., 2019. Root cause detection with an ensemble machine learning approach in the multivariate manufacturing process. Industr. Eng. Big Data Era, 163–174. https://doi.org/10.1007/978-3-030-03317-0_14.

Drotár, P., Gazda, J., Smékal, Z., 2015. An experimental comparison of feature selection methods on two-class biomedical datasets. Comput. Biol. Med. 66, 1–10. https://doi.org/10.1016/j.compbiomed.2015.08.010.

Du J, Jin W, Cai Z, Zhu F, Wu Z, Lu H, editors. A New Feature Evaluation Algorithm and Its Application to Fault of High-Speed Railway. In: Proceedings of the Second International Conference on Intelligent Transportation. ICIT 2016. Smart Innovation, Systems and Technologies: 2016 Oct 25; Singapore, Springer; 2016.

Dunne, K., Cunningham, P., Azuaje, F., 2002. Solutions to Instability Problems with Sequential Wrapper-Based Approaches to Feature Selection. Tech rep, Trinity College.

Fernandez-Lozano, C., Seoane, J.A., Gestal, M., et al., 2015. Texture classification using feature selection and kernel-based techniques. Soft Comput. 19 (9), 2469–2480. https://doi.org/10.1007/s00500-014-1573-5.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural. Comput. 4 (1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1.

George, G.V.S., Cyril Raj, V.C., 2015. Accurate and stable feature selection powered by iterative backward selection and cumulative ranking score of features. Indian J. Sci. Technol. 8 (11). https://doi.org/10.17485/ijst/2015/v8i11/71766.

Gimenez, J.R., Zou, J., 2018.. Improving the stability of the knockoff procedure: multiple simultaneous knockoffs and entropy maximization. CoRR abs/1810.11378.

Ginsburg, S.B., Lee, G., Ali, S., Madabhushi, A., 2016. Feature importance in nonlinear embeddings (FINE): applications in digital pathology. IEEE Trans. Med. Imag. 35 (1), 76–88. https://doi.org/10.1109/TMI.2015.2456188.

Goh, W.W.B., Lee, Y.H., Ramdzan, Z.M., et al., 2012. Proteomics signature profiling (PSP): a novel contextualization approach for cancer proteomics. J. Proteome Res. (11), 1571–1581 https://doi.org/10.1021/pr200698c.

Goh, W.W.B., Wong, L., 2016. Evaluating feature-selection stability in next-generation proteomics. J. Bioinform. Comput. Biol. 14 (5). https://doi.org/10.1142/S0219720016500293.

Grollemund, P.M., Abraham, C., Baragatti, M., Pudlo, P., 2019. Bayesian functional linear regression with sparse step functions. Bayesian Anal. 14, 111–135. https://doi.org/10.1214/18-BA1095.

Gulgezen G, Cataltepe Z, Yu L. Stable and Accurate Feature Selection. In: Proc 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I. 2009:455-468. https://doi.org/10.1007/978-3-642-04180-8_47.

Han, Y., Yu, L., 2012. A variance reduction framework for stable feature selection. Stat. Anal. Data Min. 5 (5), 428–445. https://doi.org/10.1002/sam.11152.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning. Springer, New York.

Haury, A.-C., Gestraud, P., Vert, J.P., 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One 6, (12). https://doi.org/10.1371/journal.pone.0028210 e28210.

He, Z., Yu, W., 2010. Stable feature selection for biomarker discovery. Comput. Biol. Chem. 34, 215–225. https://doi.org/10.1016/j.compbiolchem.2010.07.002.

He, W., Zhu, X., Cheng, D., et al., 2016. Low-rank unsupervised graph feature selection via feature self-representation. Multimed. Tools Appl. 76, 12149–12164. https://doi.org/10.1007/s11042-016-3937-6.

Hinrichs, A., Prochno, J., Ullrich, M., 2019. The curse of dimensionality for numerical integration on general domains. J. Complex. 50, 25–42. https://doi.org/10.1016/j.jco.2018.08.003.

Hua, R., Zhu, X., Cheng, D., et al., 2017. Graph self-representation method for unsupervised feature selection. Neurocomputing 220, 130–137. https://doi.org/10.1016/j.neucom.2016.05.081.

Huang, G.T., Cunningham, K.I., Benos, P.V., Chennubhotla, C.S., 2013. Spectral clustering strategies for heterogeneous disease expression data. Pac. Symp. Biocomput., 212–223

Huang, G.T., Tsamardinos, I., Raghu, V., Kaminski, N., Benos, P.V., 2015. T-RECS: stable selection of dynamically formed groups of features with application to prediction of clinical outcomes. Pac. Symp. Biocomput. 20, 431–442.

Isachenko, R.V., Strijov, V.V., 2018. Quadratic programming optimization with feature selection for nonlinear models. Lobachevskii J. Math. 39 (9), 1179–1187. https://doi.org/10.1134/S199508021809010X.

Jacob, L., Obozinski, G., Vert, J.P., 2009. Group lasso with overlap and graph lasso. In: Proc 26th international conference on machine learning. ACM, pp. 433–440.

Jeitziner, R., Carrière, M., Rougemont, J., Oudot, S., Hess, K., Brisken, C., 2019. Two-Tier Mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz052.

Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. 12 (1), 95–116. https://doi.org/10.1007/s10115-006-0040-8.

Kamkar, I., 2016. Building stable predictive models for healthcare applications: a data-driven approach [dissertation], School of Information Technology. Deakin Univeristy.

Kamkar, I., Gupta, S.K., Phung, D., Venkatesh, S., 2015. Stable feature selection for clinical prediction: exploiting ICD tree structure using Tree-Lasso. J. Biomed. Inform. 53, 277–290. https://doi.org/10.1016/j.jbi.2014.11.013.

Kamkar, I., Gupta, S.K., Phung, D., Venkatesh, S., 2015. Stable feature selection with support vector machines. In: Australasian Joint Conference on Artificial Intelligence, pp. 298–308. https://doi.org/10.1007/978-3-319-26350-2_26.

Kamkar, I., Gupta, S.K., Phung, D., Venkatesh, S., 2015. Exploiting Feature Relationships Towards Stable Feature Selection. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA); Paris, pp. 1–10. 10.1109/DSAA.2015.7344859.

Kanal, L., Chandraskekaran, B., 1971. On dimensionality and sample size in statistical pattern classification. Pattern Recognit. 3, 225–234.

Kang, C., Huo, Y., et al., 2019. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. J. Theor. Biol. 463, 77–91. https://doi.org/10.1016/j.jtbi.2018.12.010.

Katrutsa, A., Strijov, V., 2017. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Syst. Appl. 76, 1–11.

Khoshgoftaar TM, Fazelpour A, Wan H, Wald R. A Survey of Stability Analysis of Feature Subset Selection Techniques. In: IEEE 14th International Conference on Information Reuse & Integration (IRI); San Francisco, CA.; 2013. p.424-431. 10.1109/IRI.2013.6642502.

Kumar, V., Minz, S., 2014. Feature selection: a literature review. Smart Comput Rev. 4 (3), 211–229. https://doi.org/10.6029/smartcr.2014.03.007.

Kumar, A.P., Valsala, P., 2013. Feature selection for high dimensional DNA microarray data using hybrid approaches. Bioinform. 9 (16), 824–828. https://doi.org/10.6026/97320630009824.

Kuncheva, L.I., 2007. A stability index for feature selection. 25th Multi-Conference. on Applied Informatics, IASTED International Conference on Artificial Intelligence and Applications.

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., Tsamardino, I., 2017. Feature selection with the R package MXM: discovering statistically equivalent feature subsets. J. Statistical. Softw. 80 (7). https://doi.org/10.18637/jss.v080.i07.

Lahmiri, S., Shmuel, A., 2019;March(49):. Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine. Biomed. Signal Process Control. 49, 427–433. https://doi.org/10.1016/j.bspc.2018.08.029.

Lan, Z., 2018. EEG-based emotion recognition using machine learning techniques (Doctoral dissertation). Nanyang Technological University, Singapore. http://hdl.handle.net/10220/46340.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M.A., Huang, F.J., 2006. A Tutorial on Energy-Based Model. Predicting Structured Data. MIT Press, Cambridge, MA, USA.

Li, Y., Si, J., Zhou, G., Huang, S., Chen, S.F.R.E.L., 2015. A stable feature selection algorithm. IEEE Trans Neural Netw Learn Syst. 26 (7). https://doi.org/10.1109/TNNLS.2014.2341627.

Li, Y., Li, T., Liu, H., 2017. Recent advances in feature selection and its applications. Knowl. Inf. Syst. 53 (3), 551–577. https://doi.org/10.1007/s10115-017-1059-8.

Lim, K., Wong, L., 2014. Finding consistent disease subnetworks using PFSNet. Bioinformatics 30 (2), 189–196. https://doi.org/10.1093/bioinformatics/btt625.

Liu, Y., Diao, X., Cao, J., Zhang, L., 2017. Evolutionary Algorithms' Feature Selection Stability Improvement. In: System. International Conference on Bio-Inspired Computing: Theories and Applications, pp. 68–81. https://doi.org/10.1007/978-981-10-7179-9_6.

Liu, Z., Wang, R., Japkowicz, N., et al., 2019. Mobile app traffic flow feature extraction and selection for improving classification robustness. J. Netw. Comput. Appl. 125, 190–208. https://doi.org/10.1016/j.jnca.2018.10.018.

Loscalzo, S., Yu, L., Ding, C., 2009. Consensus group stable feature selection. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Paris, France. New York, pp. 567–576. Doi: 10.1145/1557019.1557084.

Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S., 2009. Measuring stability of feature selection in biomedical datasets. AMIA Annu. Symp. Proc., 406–410

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intra-class correlation coefficients. Psychol. Methods 1 (1), 30–46.

Meinshausen, N., Buhlmann, P., 2010. Stability selection. J. R. Stat. Soc. 72 (4), 417–473.

Mirjalili, S., Mirjalili, S.M., Lewis, A., 2014. Grey wolf optimiser. Adv. Eng. Softw. 69, 46–56. https://doi.org/10.1016/j.advengsoft.2013.12.007.

Mohammadi, M., Noghabi, H.S., Hodtani, G.A., Mashhadi, H.R., 2016. Robust and stable gene selection via maximum-minimum correntropy criterion. Genom. 107, 83–87. https://doi.org/10.1016/j.ygeno.2015.12.006.

Mohana, C.P., Perumal, K., 2016. A survey on feature selection stability measures. Int. J. Comput. Inf. Technol. 5 (1).

Moon, M., Nakai, K., 2016. Stable feature selection based on the ensemble L1-norm support vector machine for biomarker discovery. BMC Genom. 17 (13), 1026. https://doi.org/10.1186/s12864-016-3320-z.

Mostafa, S.A., Mustapha, A., Mohammed, M.A., Hamed, R.I., Arunkumar, N., Ghani, M.K.A., et al., 2019. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. Cogn. Syst. Res. 54, 90–99. https://doi.org/10.1016/j.cogsys.2018.12.004.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012

Nogueira, S., 2018. Quantifying the stability of feature selection (Doctoral dissertation). University of Manchester, United Kingdom.

Nogueira, S., Brown, G., 2016. Measuring the Stability of Feature Selection. Joint Eur. Conf. Mach. Learn. Knowledge Discov. Databases, 442–457. https://doi.org/10.1007/978-3-319-46227-1_28.

Park, C.H., Kim, S.B., 2015. Sequential random k-nearest neighbor feature selection for high-dimensional data. Expert. Syst. Appl. 42, 2336–2342. https://doi.org/10.1016/j.eswa.2014.10.044.

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., Aerts, H.J.W.L., 2015. Machine learning methods for quantitative radiomic biomarkers. Sci. Rep. 5, 13087. https://doi.org/10.1038/srep13087.

Patil, M., Rao, M., 2018. Studying the contribution of machine learning and artificial intelligence in the interface design of e-commerce site. Smart Intell. Comput. Appl., 197–206 https://doi.org/10.1007/978-981-13-1927-3_20.

Perthame, E., Friguet, C., Causeur, D., 2016. Stability of feature selection in classification issues for high-dimensional correlated data. Stat. Comput. 26, 783–796. https://doi.org/10.1007/s11222-015-9569-2.

Ramondta, S., Ramírezb, A.S., 2019. Assessing the impact of the public nutrition information environment: adapting the cancer information overload scale to measure diet information overload. Patient Educ. Couns. 102, 37–42. https://doi.org/10.1016/j.pec.2018.07.020.

Randall, R.B., Antoni, J., Smith, W.A., 2019. A survey of the application of the cepstrum to structural modal analysis. Mech. Syst. Signal. Process. 118, 716–741. https://doi.org/10.1016/j.ymssp.2018.08.059.

Rondina, J.M., Hahn, T., et al., 2014. SCoRS – a method based on stability for feature selection and apping in neuroimaging. IEEE Trans. Med. Imag. 33 (1). https://doi.org/10.1109/TMI.2013.2281398.

Sakae, Y., Straub, J.E., Okamoto, Y., 2018. Enhanced sampling method in molecular simulations using genetic algorithm for biomolecular systems. J. Comput. Chem. https://doi.org/10.1002/jcc.25735.

Sánchez, L.E., Diaz-Pace, J.A., Zunino, A., 2019. A family of heuristic search algorithms for feature model optimization. Sci. Comput. Progr 172, 264–293. https://doi.org/10.1016/j.scico.2018.12.002.

Selvaraj, G., Kaliamurthi, S., Kaushik, A.C., Khan, A., Wei, Y.K., Cho, W.C., et al., 2018. Identification of target gene and prognostic evaluation for lung adenocarcinoma using gene expression meta-analysis, network analysis and neural network algorithms. J. Biomed. Inf. 86, 120–134. https://doi.org/10.1016/j.jbi.2018.09.004.

Sharma, P., Gupta, A., et al., 2018. The health of things for classification of protein structure using improved grey wolf optimization. J. Supercomput., 1–16 https://doi.org/10.1007/s11227-018-2639-4.

Soh, Donny, Dong, Difeng, Guo, Yike, Wong, Limsoon, 2011. Finding consistent disease subnetworks across microarray datasets. BMC Bioinform. 12 (S13). https://doi.org/10.1186/1471-2105-12-S13-S15.

Somol, P., Novovicová, J., 2010. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Trans. Pattern. Anal. Mach. Intell. 32 (11), 1921–1939. https://doi.org/10.1109/TPAMI.2010.34.

Storn, R., Price, K., 1997. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. 11 (4), 341–359. https://doi.org/10.1023/A:1008202821328.

Taylor, A., Steinberg, J., Andrews, T.S., Webber, C., 2015. GeneNet Toolbox for MATLAB: a flexible platform for the analysis of gene connectivity in biological networks. Bioinform. 31 (3), 442–444. https://doi.org/10.1093/bioinformatics/btu669.

Tibshirani, R., Saunders, M., et al., 2005. Sparsity and smoothness via the fused lasso. J. R. Statist. Soc. B. 67 (1), 91–108.

Tichý, O., Bódiová, L., Šmídl, V., 2019. Bayesian non-negative matrix factorization with adaptive sparsity and smoothness prior. IEEE Signal. Process Lett. 26 (3), 510–514. https://doi.org/10.1109/LSP.2019.2897230.

Tran, L., Kossaifi, J., Panagakis, Y., et al., 2019. Disentangling geometry and appearance with regularised geometry-aware generative adversarial networks. Int. J. Comput. Vis. 127 (6–7), 824–844. https://doi.org/10.1007/s11263-019-01155-7.

van Erp, M., Schomaker, L., 2000. Variants of the Borda count method for combining ranked classifier hypotheses. In: Proceedings 7th International Workshop on frontiers in handwriting recognition (7th IWFHR), pp. 443–452.

Villaruz, L.C., Huang, G., Romkes, M., Kirkwood, J.M., Buch, S.C., Nukui, T., et al., 2015. MicroRNA expression profiling predicts clinical outcome of carboplatin/paclitaxel-based therapy in metastatic melanoma treated on the ECOG-ACRIN trial E2603. Clin. Epigenet. 7 (1). https://doi.org/10.1186/s13148-015-0092-2.

Wade, B.S.C., Joshi, S.H., Gutman, B.A., Thompson, P.M., 2017. Machine learning on high dimensional shape data from subcortical brain surfaces: a comparison of feature selection and classification methods. Pattern Recognit. 63, 731–739. https://doi.org/10.1016/j.patcog.2016.09.034.

Wan, C., 2019. Feature Selection Paradigms. In: Hierarchical Feature Selection for Knowledge Discovery. Advanced Information and Knowledge Processing. Springer, Cham, pp. 17–23. https://doi.org/10.1007/978-3-319-97919-9_3.

Wang, B., Chiang, H.D., 2011. ELITE: ensemble of Optimal Input-Pruned Neural Networks Using TRUST-TECH. IEEE Trans. Neural. Netw. 22 (1). https://doi.org/10.1109/TNN.2010.2087354.

Xiao, H., Biggio, B., et al., 2015. Is Feature Selection Secure against Training Data Poisoning? Proceedings of the 32 nd International Conference on Machine Learning, Lille, France.

Xin, B., Hu, L., Wang, Y., Stable, Gao W., 2015. Feature Selection from Brain sMRI. Proc Twenty-Ninth AAAI Conference on Artificial Intelligence.

Xu, H., Caramanis, C., Mannor, S., 2012. Sparse algorithms are not stable: a no-free-lunch theorem. IEEE Trans. Pattern. Anal. Mach. Intell. 34 (1), 187–193. https://doi.org/10.1109/TPAMI.2011.177.

Yan, K., Zhang, D., 2012. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sens. Actuators B Chem. 212, 353–363. https://doi.org/10.1016/j.snb.2015.02.025.

Yang, F., Mao, K.Z., 2011. Robust feature selection for microarray data based on multi-criterion fusion. IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (4), 1080–1092. https://doi.org/10.1109/TCBB.2010.103.

Ye, G.B., Chen, Y., Xie, X., 2011. Efficient variable selection in support vector machines via the alternating direction method of multipliers. In: International Conference on Artificial Intelligence and Statistics, pp. 832–840.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J R Statist. Soc. B. 68 (1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x.

Zheng, W., Eilamstock, T., Wu, T., Spagna, A., et al., 2019. Multi-feature based network revealing the structural abnormalities in autism spectrum disorder. IEEE Trans. Affective Comput. 1 (1). https://doi.org/10.1109/TAFFC.2018.2890597.

Zhu, X., Huang, Z., Cheng, H., Cui, J., Shen, H.T., 2013. Sparse hashing for fast multimedia search. ACM Trans Inf Syst. 31 (2), 9.1–9.24.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy Stat. Soc. Ser. B (Stat Methodol). 67, 301–320.

Please cite this article as: U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.06.012