Stability of Feature Selection Algorithms

Dipti Theng and K. K. Bhoyar

Abstract Feature selection is in great demand as it is a crucial phase in classification, clustering, and prediction tasks all around the different sectors. The stability of the feature selection process determines the sensitivity of the selection to dataset variation. In bioinformatics, a task of biomarker identification requires selection of disease associated genes (relevant feature subset), to be insensitive to the variations in the training set, which can give confidence to the domain experts and can prevent result manipulation in the real-world applications. Traditional feature selection strategies are ineffective to capture this high sensitivity to the numerous perturbations. This motivates the need of robust feature selection strategy which produces stable feature set across datasets from multiple domains.

Keywords Machine learning · Feature selection · Stability · Stability measures

1 Introduction

A feature is an attribute or variable used to characterize some property of individual data objects in data mining, machine learning, and data analytics [1]. In high-dimensional datasets, each individual pattern vector can contain tens to hundreds, or even thousands, of features. Recognition tasks in biology, diagnostics tasks, and biomarkers identification in health care and genetic engineering, character, text, and face recognition from digital images, spam email identification, astronomy, and economics are all examples of such data. Feature selection (FS) is a challenging topic in pattern recognition with many potential applications. The number of samples required for every dataset grows exponentially in proportion to the dimensionality of the feature space [1]. Thus, huge data storage and computational resources are required for such datasets. The curse of dimensionality is the term coined to this

D. Theng (✉) · K. K. Bhoyar
Department of Computer Technology, YCCE, Nagpur, India
e-mail: deepti.theng@gmail.com

M. Gupta et al. (eds.), *Artificial Intelligence on Medical Data*,
Lecture Notes in Computational Vision and Biomechanics 37,
https://doi.org/10.1007/978-981-19-0151-5_26

phenomenon [2]. Reducing the dimensionality of the input feature space is a prominent solution to this problem and to employ only those features that are most relevant to the given problem. There are two types of feature reduction:

Feature Selection: It is the process of selecting a minimal number of features from a vast range of features. Because of the smaller feature set, certain techniques are computationally possible to implement.

Feature Extraction: The goal of feature extraction is to create new features that are not always the consequence of feature transformations. For example, the features defined from patterns are generated features.

One successful method for identifying important features for dimensionality reduction is feature selection. It is a key requirement of every data analytics or machine learning task. Informative features are the useful features for describing the underlying objectives. They are vital for producing accurate and easy to explain predictive models and yield good results in various data analytics part [1]. Feature selection is a method for selecting the best subset of features based on a set of criteria. Feature selection is needed:

- To reduce the size of feature set.
- To evaluate the usefulness of features and feature set.
- To improve model performance in terms of predictive power, speed, and model simplicity.
- To assist in model selection by visualizing data.

The stability of the feature selection process determines the sensitivity of the selection to dataset variation. The stability of the results can be determined by comparing them pair by pair. The stability is higher when the similarity is larger. Stability of feature selection methods is an important criterion indicating how sensitive the solution (i.e., selected features) is to agitate in the input training data. The output of the feature selection algorithms can be represented in three ways: indexing, ranking, and weighting. The stability of the feature selection is a relatively ignored concern in work on high-dimensional problems and in general for tasks requiring feature selection. Moreover, the manuscript will concentrate on the stability and general design of feature selection method and on the evaluation methodology applied in feature selection.

1.1 Feature Selection from High-Dimensional Data

Major emerging areas like IoT, AI, and Big Data have tremendously increased data size and dimensionality of the data. As pointed by the curse of dimensionality, for any dataset with the size of the feature space, the number of samples required grows exponentially. It has fetched many challenges for traditional approaches in the area of machine learning and data analytics. This has fetched an attention of researchers to need to work on dimensionality reduction for the high-dimensional data. Another motive behind high-dimensional data future set reduction is to reduce the computational overhead. Processing massive amount of data for high dimensionality (large number of observed features) consumes more resources and time as reported by many literatures [3–8].

1.2 Need of Stable Feature Selection

Stability of feature is defined as the selection of most robust feature set which does not change with data or function perturbations. Recent research on feature selection has gained an attention on stability of features a major claim by the bioinformatics and biomedical fields. Feature selection is in great demand as it is a crucial phase in classification, clustering, and prediction tasks all around the different sectors including finance, health care, biomedical, industry, and social media. Such diversified application area for feature selection requires an approach which is independent of the classifier type, feature selection approach, and unbiased to the dataset. A rational of 'No Free Lunch Theorem' no algorithm that can outperform stochastic search on all problems can be applied to feature selection. Goodness of an algorithm for feature selection is defined by the way it address the optimal trade-off between accuracy and stability [9]. Past work presented in [10] has analyzed that the existing algorithms are deficient to evaluate joint performance to stability and accuracy on the dataset at hand. Stable feature selection has been emerged as a strong evidence for disease associated biomarker selection (stable feature selection) to bioinformatics problems of microarray and mass spectrometry datasets [11].

In bioinformatics, a task of biomarker identification requires selection of disease associated genes (relevant feature subset), to be insensitive to the variations in the training set, which can give confidence to the domain experts and can prevent result manipulation in the real-world applications [12]. Traditional feature selection strategies are ineffective to capture this high sensitivity to the numerous perturbations. This motivates the need of robust feature selection strategy which produces stable feature set across datasets from multiple domains.

1.3 Factors for Instability: Why Instability May Occur?

Major cause of instability in many feature selection algorithm is that, they are classifier dependent (targeting to achieve highest accuracy) and do not consider stability into account, thus, causing limited reproducibility of the result. Presence of high noise, variance, and outlier samples in the dataset may cause instability in the feature selection. Another major cause of instability in bioinformatics applications is due to high dimensional with low sample size dataset which is more susceptible to the irrelevant and redundant feature inclusion in the feature subset generation [13]. Apart from the insensitivity to the change in training dataset (data perturbation), the stability is also highly dependent on the choice of feature selection approach (i.e., function perturbation) considered for a task in hand. Beside of small sample size, other causes of instability such as type of feature selection algorithm and different partitioning of the training–testing sets have also been proved in [14]. An algorithm may hold one or more causes of instability.

In ensemble selection methods, the choice of ensemble parameters like, aggregation function, ensemble size, method selected in ensemble, etc. may critically affect the feature selection performance and stability of the feature selection [15]. Also, feature selection algorithm stability is significantly affected by the dataset characteristics. Authors in [2] have presented comprehensive analysis on dependency of feature selection stability to dataset characteristics, which take account of feature set dimensionality, sample size, and variation of data distribution across folds. Many research studies claimed that, the feature selection outcome (feature subset) stability has a strong dependency on the size of sample. In [16], authors have discussed that the dataset with large number of redundant features is more prone to the unstable feature selection. Removal of redundant features is necessary in order to obtain stable feature set. Several techniques that gained popularity in improving feature selection stability are: prior feature relevance (domain or expert knowledge), ensemble feature selection, group feature selection, and sample injection.

2 Literature Review

2.1 Stability Measures

In the similarity-based method, Dunne et al. looked at the idea in the context of wrapper-based feature selection, defining stability as the average pairwise similarity between the $M(M-1)$ possible pairs of feature sets. The choice of a similarity measure has a significant impact on the Dunne definition for stability measure and its properties [17]. Regardless of whether feature selection approach was employed, Kalousis et al. were among the first to discuss stability in depth. Authors have not addressed how best to quantify stability [3]. Yu et al. introduced a stability metric based on the Dice similarity measure, which shows that stability grows predictably with the number of features picked, favoring bigger feature sets [4].

Zucknick et al. create a correlation-accounting form of the Jaccard index. It is not possible to compare feature sets of various sizes [5]. Although Shi et al. used the Percentage of Overlapping Genes (POG) index, the measured percentage of overlapping genes (POG) of DEG lists from two research is most likely to be low. It is used only in the case of no redundancy [6]. Kuncheva's stability index does not make use of correlation coefficients of sampling experiments. Instead, all subsampling experiments are combined using a simple accumulation of feature subsets [7].

Lustgarten et al. suggested a modified version of Kuncheva's index that allows feature sets of various sizes to be compared. It is ineffective to satisfy the property of correction-for-chance [8]. Wald et al. projected a stability measure based on similarity-based measure. Constants are not a constraint on Wald's stability measure, as it has been observed [9]. Zhang et al. devise a variant of the POG as normalized POG (nPOG). The stability measurements based on POG and nPOG similarity measures have been found to be unbounded [15]. Goh and Wong formulated a frequency-based stability measure. Goh's measure violates the backward implication. It employs variable stability values that are determined by the number of features selected (k) [18]. Davis et al. penalize frequency in order to account the increase in stability by artificial which tends to occur with gene signatures get longer. Monotonicity is a quality that this metric lacks [19].

Each conceivable feature set of k features is treated as a random variable by Krizek et al., who estimate its Shannon entropy. This measure records lower values corresponding to higher stability [20]. Guzman et al. defined feature sets stability measure of fixed cardinality k. Guzman's stability measure produces weaker stability as well as a smaller average variance, violating the monotonicity requirement [13]. Somol et al. suggested CWrel metric which mandates a stability measure that is constrained by constants. This measure violates the forward implication and thus fails the property of correction-for-chance [21]. Lausser et al. proposed a metric for feature sets of k dimensional. Lausser's measure does not satisfy the property 'fully defined' because they are only defined when the quantity of features chosen is predetermined [22].

The researchers have proposed new metrics for measuring stability, experimental designs for testing the stability of feature selection, and new strategies to increase feature selection stability, as described above. This section has briefly analyzed the gap in present work on the feature selection stability.

Many researchers have studied an effect and applications of the stability on feature selection. Table 2 summarized the important finding by these authors for the stability of feature selection. It has listed important finding by the researchers for the effect of stability on feature selection. From these findings, it is observed that stable feature selection improves the performance and confidence of the feature subset selection for various applications. Major application of the stable feature selection is in the area of bioinformatics where stable feature subset represents the true biomarkers.

3 Stable Feature Selection Algorithms

Most machine learning and data analytics techniques found not effective for high-dimensional data. It has been observed that some approaches perform well for some datasets but not for others. This is mostly due to the dataset characteristics. As a result, studying the features of the dataset is critical for recommending a suitable feature selection algorithm. Review of the work done on stability index is designed considering the characteristics of specific dataset and its applications. No single

Table 1 Stability measures

S. No.	Type of measure	Name	First used in	Measure	[min; max]	Fully defined	Monotonicity	Bounds	Maximum	Correction
1	Similarity-based approach	Hamming	Dunne et al. (2002)	$1 - \frac{\lvert s_i \setminus s_j \rvert + \lvert s_i \setminus s_j \rvert}{d}$	[0; 1]	√	√	√	√	
2		Jaccard	Kalousis et al. (2005)	$\frac{r_{i,j}}{\lvert s_i \cup s_j \rvert}$	[0; 1]	√	√	√	√	
3		Dice	Yu et al. (2008)	$\frac{2r_{i,j}}{\lvert k_i + k_j \rvert}$	[0; 1]	√	√	√	√	
4		Ochiai	Zucknick et al. (2008)	$\frac{r_{i,j}}{\sqrt{k_i k_j}}$	[0; 1]	√	√	√	√	
5		POG	Shi et al. (2006)	$\frac{r_{i,j}}{k_i}$	[0; 1]	√	√	√	√	
6		Consistency	Kuncheva (2007)	$\frac{r_{i,j} - \frac{k^2}{d}}{k - \frac{k^2}{d}}$	[-1; 1]		√	√	√	√
7		Lustgarten	Lustgarten et al. (2009)	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{\min(k_i, k_j) - max(0, k_i + k_j - d)}$	[−1; 1]	√	√	√		√
8		Wald	Wald et al. (2013)	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{\min(k_i, k_j) - \frac{k_i k_j}{d}}$	[1 − d; 1]	√	√			√
9		nPOG	Zhang et al. (2009)	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{k_i - \frac{k_i k_j}{d}}$	[1 − d; 1]	√	√		√	√

(continued)

Table 1 (continued)

S. No.	Type of measure	Name	First used in	Measure	[min; max]	Fully defined	Monotonicity	Bounds	Maximum	Correction
10	Frequency-based approach	Goh	Goh and Wong (2016)	$\hat{\Phi}_{\mathrm{Goh}}(\mathcal{Z}) = \frac{1}{d}\sum_{f=1}^{d} p_f$	[0; 1]	√		√		
11		Davis	Davis et al. (2006)	$\hat{\Phi}_{\mathrm{Davis}}(\mathcal{Z}) =$ $\max\left(0, \frac{1}{f}\sum_{f=1}^{d} \hat{p}_f - \alpha\frac{\mathrm{median}(k_1, k_2, \ldots, k_M)}{d}\right)$	[0; 1]	√		√		
12		Krizek	Krizek et al. (2007)	$\hat{\Phi}_{\mathrm{Krizek}}(\mathcal{Z}) = -\sum_{s_i \in \mathcal{Z}} \hat{p}(s_i)\log_2 \hat{p}(s_i)$	$\left[0, \log\left(min\left(M, \binom{d}{k}\right)\right)\right]$				√	
13		Guzman	Guzman et al. (2011)	$\hat{\Phi}(\mathcal{Z}) = 1 - \frac{D_{\mathrm{JS}}(q_1, \ldots, q_M)}{D^*_{\mathrm{JS}}(q_1, \ldots, q_M)}$	[0; 1]			√	√	√
14		CWrel	Somol et al. (2010)	$\hat{\Phi}(\mathcal{Z}) =$ $\frac{d\left(M\bar{k} - D + \sum_{f=1}^{d} M\hat{p}_f(M\hat{p}_f - 1)\right) - (M\bar{k})^2 + D^2}{d(H^2 + M(M\bar{k} - H) - D) - (M\bar{k})^2 + D^2}$	[0; 1]	√	√	√		
15		Lausser	Lausser et al. (2013)	$\hat{\Phi}(\mathcal{Z}) = \frac{1}{M^2 k}\sum_{i=1}^{M} i^2 a^{(i)}$	[0; 1]		√	√	√	

measure exists claiming to work on multiple datasets of varying characteristics. It is required to study trade-off between dataset and feature selection algorithm for suitable stability measure selection. Designing such a measure is a challenging task. Furthermore, it has been observed that feature selection methods are not wisely selected for stability analysis. The feature selection methods need to be studied for classifier-independent and classifier-dependent categories.

Table 2 Review of past stability research

S. No.	References	Major area/keywords	Important findings
1	[23]	Feature subset selection, filter, wrapper, hybrid methods, multi-objective optimization, particle swarm optimization	According to research, using a heuristic, evolutionary FSS technique repeatedly fails to yield consistent outcomes. Proposed a new hybrid (wrapper-filter) FSS technique, COMB-PSO-LS that selects the least dependent and most relevant feature subsets by combining particle swarm optimization (PSO) with a local search strategy. COMB-PSO-LS outperforms typical PSO approaches in terms of providing gene subsets that are stable and non-redundant relevant to the classification procedure, according to research
2	[24]	Feature selection, stability selection, sample-specific analysis, random lasso, L1-type regularization	A new technique termed robust sample-specific stability selection is proposed. A sample-specific random lasso based on kernel-based L1-type regularization and weighted random sampling is also proposed. According to major study, the proposed methodology performs sample-specific analysis efficiently and precisely, yielding biologically meaningful gene selection results
3	[25]	Feature selection, ensemble methods, classification, stability, high dimensionality	The reliability of feature selectors is being used to formulate an ensemble feature selection strategy. Its goal is to provide a distinctive and reliable feature selection while also considering forecasting accuracy. According to a major research finding, homogeneous ensembles constructed with unstable base learners are superior to heterogeneous ensembles for enhancing stability because they produce optimal stability results

(continued)

Table 2 (continued)

S. No.	References	Major area/keywords	Important findings
4	[26]	Dimensionality reduction, feature selection, filter method, wrapper method, gene expression data	In the research study, five feature selection techniques: relief, chi-square, information gain, random forest, and recursive feature elimination for SVM (RFE-SVM) are compared on the basis of Kuncheva index stability. It is noted that the performance of any feature selection approach on high-dimensional microarray data is dependent not only on the accuracy of the classifier, but also on the method's stability. In terms of stability, the two filter methods information gain and chi-square method beat other methods, and their stability grows with subset size
5	[27]	Stability, data selection bias, interpretability	Developed a framework for assessing the influence of deliberate stability, as well as trials that were parametric to a variety of pre-processing procedures and classification algorithms. Stability, overfitting, and accuracy are contrasting objectives, which then require a trade-off analysis. According to the findings of the study, a stability impact evaluation should be considered alongside predictive performance analysis
6	[28]	Machine learning, feature selection, stability, fMRI	A unique feature selection approach based on a single-layer neural network that integrates cross-validation in feature selection as well as stability selection via iterative sub-sampling is proposed. A comparative analysis of the proposed approach demonstrated improved classifier accuracy, lower processing costs, and more consistency in the selection of important features. Future study can focus on figuring out how to best optimize neural networks for stable detection of relevant features in a variety of datasets

(continued)

Table 2 (continued)

S. No.	References	Major area/keywords	Important findings
7	[29]	Stacking, stacked, generalization, bagging, bootstrap, algorithmic, stability, generalization	Dag-stacking and bag-stacking, the hypothesis stability of stacking is investigated by the authors. Demonstrated a relationship between weighted bagging and bag-stacking, showing that bag-stacking equals weighted bagging. Research study showed that stacking improves the stability of the stacked learning algorithms by a factor of $1/m$
8	[30]	Feature selection, evolutionary algorithm, multi-objective problem, stability ensemble	The authors presented feature selection wrapper technique based on a multi-objective evolutionary algorithm. In order to assess the stability of the suggested wrapper technique, a novel feature ranking procedure is proposed
9	[31]	Algorithmic stability, hypothesis stability, pointwise hypothesis stability, decision tree, logistic regression, stability measures	For decision trees and logistic regression, derive two ideas of stability: hypothesis and pointwise hypothesis stability. Also, presented a stability measuring framework to measure these notions of stability. The stability of decision trees has been shown to be dependent on the tree's depth, and that it is determined by the cross-entropy loss's smallest eigenvalue of the Hessian matrix. Finding of the study: The learning algorithm logistic regression is not stable because it is dependent on an uncontrollable parameter for its stability
10	[32]	Stability, generalization, stochastic gradient descent, kernel activation functions	The generalization capabilities of non-parametric activation functions, such as kernel activation functions, were investigated (KAFs). KAFs add further parameters to the learning process, allowing nonlinearities to be adjusted per neuron independently. When trained with SGD for a finite number of steps, a NN endowed with KAFs is stable and generalizes well, according to a major research conclusion

(continued)

Table 2 (continued)

S. No.	References	Major area/keywords	Important findings
11	[33]	Feature selection, stability, perturbation, knowledge discovery, robustness instability	An overview of feature selection strategies is given, as well as discussed the issue of instability of the feature selection algorithm. Few solutions for dealing with various sources of instability are presented. In the domains of health care, bioinformatics, image analysis, and business analytics, feature selection strategies and stability analysis were discussed
12	[34]	Feature selection, stability, feature correlations	Presented estimation of feature selection stability, on given correlations in the data or domain knowledge. Proposed a new measure that enables information regarding correlated and/or semantically relevant variables to be incorporated. Defined a new stability measure which satisfies these properties
13	[35]	Feature selection, wrapper method, stability, classification accuracy	Proposed a stability improved feature selection algorithm. The sequential forward selection approach is utilized in research to define the upper bound of the confidence interval for classification accuracy. The results of the experiments revealed that this approach can meet classification accuracy and stability requirements simultaneously

The most often widely used methods are group feature selection and ensemble feature selection (EFS). For relevant feature selection, these methods necessitate a thorough examination of the ideal number of components in ensembles. In the case of ensemble feature selection, a priori ensemble size determination has yet to be addressed; hence, statistical tests are currently utilized to determine the appropriate number of components. Present EFS is computationally expensive for implementation. There is a potential for optimization in the future work.

Existing literatures have addressed stability in presence of feature correlation. A similarity metric is required to analyze the overlap of two feature subsets in order to assess the stability of feature selection strategies. It requires the creation of new ways to help eliminate the redundancy that can occur when aggregating the partial feature subsets obtained from individual feature selectors. It has been noted that, feature subset is not stable with respect to the properties of stability measures. It is necessary to investigate function perturbation on each bootstrap dataset, which will result in M ranked lists for each EFS. Stability of unsupervised feature selection algorithms has not been well studied yet.

The primary key for the improved stability and performance of the solution (i.e., selected features) is a good estimator of the search criterion functional values. A better estimate means that a feature selection algorithm is more likely to converge to its (sub)optimal solution on the given problem with respect to the criterion function and the search strategy [11, 16, 23, 36]. A proper use of the available data can help to improve the reliability and performance of feature selection results to some extent.

3.1 Major Application Areas Demanding Stable Feature Selection

Different application domains obsessive of stable feature selection are: study of biomarkers in genomics, proteomics, metabolomics, and mass spectrometry data; biomedical signal processing, medical imaging; data analysis for text mining, social media, and industrial applications; audio and video analysis, etc. Each one of it demands stable attribute selection (feature selection), which helps in generating consistent results irrespective to the background noise in data. Feature dimensionality reduction in multivariate time series is explored extensively in [23].

High-dimensional disease data to identify microbial biomarkers for disease prediction, e.g., mental illness (e.g., Alzheimer, depression), migraine, inflammatory bowel disease, and type-2 diabetes [22]. Omics data such as genomics, proteomics or metabolomics is available, which provides ample evidence to study and build models for healthcare customization [37]. Study and application of machine learning in genomic microarray dataset have largely been used to determine the risk of getting cancer. Affected genes from the gene expression profile show prominent patterns linked to disease and help in effective personalized diagnosis [38]. Objective of stable feature selection to the cancer prediction is to identify highly stable and most informative features (genes) which are responsible for causing cancer [39, 40]. On cancerous RNA-Seq data, stable feature selection allows for the discovery of strand-specific information, which is a crucial element of gene regulation. The analysis of somatic clonal growth in normal tissues to determine the occurrence and extent of clonal expansion across human tissues is another application of stable feature selection. Application of stable feature selection to build a correlation between the microbiome and disease syndromes can help with personalized medication.

Feature selection has been in great demand for predicting cloud computing resource utilization and intrusion detection. Task of predicting future cloud resource usage is considered a time series analysis task. In this task, resource metrics are termed as feature set which usually includes running job count, memory usage, and cache, time of disk input/output, disk space, and CPU usage. These features are observed as independent but correlated to each other and hold cause-effect relationship [17]. Synthetic datasets are important and widely used in feature selection algorithm evaluation due to their flexibility as they provide precise information of relevant features with respect to the label/output variable [41].

3.2 Implementation of the Stability Measures

Implementation of the stability measures for the dataset and algorithm used in experimentation is quite easy as many libraries are defined for popular stability measures. Thanks to the research community making these stability measures packages available for the researchers to use and study them. Most preferred programming languages for implementation of the feature selection stability measures are; Python, MATLAB, and R-programming. These programming languages have rich packages and libraries for machine learning and stability measures, thus, making it very easy for the researchers to use and apply them to their datasets focusing more on the research objectives without worrying about the implementation part.

Some of the most popular packages for stability measures are listed in Table 3. These are the most cited packages and programming languages for the feature selection stability measures (Table 4).

4 Conclusions

An appropriate use of the available data can help to improve the reliability and performance of feature selection results to some extent. From the study presented in this manuscript, it is inferred that the existing algorithms are deficient to evaluate joint performance to stability and accuracy on the dataset at hand. Thus, there is a good future scope to apply stability for feature selection and study its effect on the accuracy. There are very few researches that presented the joint study of stability and accuracy of the feature selection. Major demanding areas for stable feature selection are bioinformatics gene expression (RNA-Seq) data analysis, biomedical image analysis, industrial optimization problems, network traffic classification, and personalized medicine, etc [42–46].

Table 3 Classes for stable feature selection approach

Class	Approach	Subtype	Description
Incorporate stability into algorithm	Ensemble feature selection	Function perturbation	It combines the results of various feature selection techniques to find a stable feature subset
		Data perturbation	Feature selectors are used in the first approach using distinct sample subsets selected from the same dataset applying random sampling, such as bagging and boosting
			The second method employs adverse feature subspaces, such as random subspace
	Prior feature relevance	Prior knowledge of the dataset is being used for the feature selection process in this approach, implying that some features are more significant than others. For instance, transfer learning	
Handling data with highly correlated features	Group feature selection	Knowledge-driven group formation	Methods that use pathway information can be used to identify groupings of related features
		Data-driven group formation	The feature selection process in this approach groups the features into clusters of similar density
Address curse of dimensionality	Sample injection	Transductive learning	In transductive learning, test data is combined with training data for feature selection
		Artificial training sample	On the basis of the available training data distribution, an artificial training sample is generated

Table 4 Implementation setup and programming languages used

S. No.	Refs.	Stability/performance measures	Experimental setup environment
1	3	Executed pairwise Spearman's Rho to calculate stability	Implemented in Python
2	6	Jaccard similarity is used to measure stability	Python, developed Asaph: an open-source toolkit for variant analysis
3	10	Implemented Kuncheva index for stability measure	Used Weka machine learning library
4	15	Implemented Jaccard index, Hamming distance, Kuncheva index	MATLAB, Weka, R, KEEL, RapidMiner, Scikit-learn, Apache Spark MLlib
5	24	Proposed a robust sample-specific stability selection method	Monte Carlo simulations
6	25	Proposed and implemented proper consistency index with pairwise comparisons	Weka machine learning workbench
7	26	Measure stability by Kuncheva index stability measure	R 3.5.1
8	27	Jaccard coefficient, Pearson's correlation coefficient	Python
9	28	SNR of the relevant features, Jaccard index, mean accuracy, standard deviation (STD)	Experiments are carried using Python version 2.7.6
10	43	The 20 stability measures were implemented in an open-source manner: Davis, Dice, Hamming, IntersectionCount, IntersectionGreedy, IntersectionMBM, IntersectionMean Jaccard Kappa, Lustgarten, Nogueira, Novovicova, Ochiai, Phi, Sechidis, Somol, Unadjusted, Wald, Zucknick, and Yu stability index	Implemented 'stabm' package, as an R (R Core Team, 2020) package that allows you to quantify the similarity of two or more feature sets
11	44	Similarity-based measure using the Jaccard index, the cosine similarity, and Lin's method	Defined a package GSimPy in Python. https://github.com/curlyae1995/GSimPy
12	45	Kuncheva index (KI), weighted consistency index (WCI), weighted aggregation: ensemble-weighted mea, Ensemble-weighted stability (top s) selection, ensemble-weighted-exponential, ensemble mean, ensemble stability (top s) selection	Used R: h2o library for LASSO, ElastNet. DNN and XGBoost library for GBM
13	46	Proposed a length adjusted stability stab	Implemented in R-programming, source code: https://github.com/zhxiaokang/EFSIS

(continued)

Table 4 (continued)

S. No.	Refs.	Stability/performance measures	Experimental setup environment
14	47	Proposed new stability measure satisfying the properties defined by the authors for stability measure	Implemented it in R, Python, and MATLAB. Implementation can be accessed at https://github.com/nogueirs/JMLR2018/tree/master/python/

References

1. Dong G, Liu H (2018) Feature engineering for machine learning and data analytics. CRC Press
2. Goswami S, Chakrabarti A, Chakraborty B (2016) A proposal for recommendation of feature selection algorithm based on data set characteristics. J UCS 22(6):760–781
3. Nogueira S, Sechidis K, Brown G (2017) On the use of Spearman's rho to measure the stability of feature rankings. In: Iberian conference on pattern recognition and image analysis. Springer, Cham
4. Chelvan P, Perumal K (2017) A comparative analysis of feature selection stability measures. In: 2017 international conference on trends in electronics and informatics (ICEI). IEEE
5. Chelvan PM, Perumal K (2017) The effects of change in statistical properties of datasets on feature selection stability. In: 2017 international conference on information communication and embedded systems (ICICES). IEEE
6. Nowling RJ, Emrich SJ (2017) Stable feature ranking with logistic regression ensembles. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE
7. Liu T et al (2017) Algorithmic stability and hypothesis complexity. In: Proceedings of the 34th international conference on machine learning, vol 70. JMLR. org
8. Zomaya AY (2013) Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. In: Biological knowledge discovery handbook: preprocessing, mining and postprocessing of biological data, vol 23, p 333
9. Brown G et al (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13:27–66
10. Pes B (2017) Feature selection for high-dimensional data: the issue of stability. In: 2017 IEEE 26th international conference on enabling technologies: infrastructure for collaborative enterprises (WETICE). IEEE
11. Ahmed S et al (2014) Multiple feature construction for effective biomarker identification and classification using genetic programming. In: Proceedings of the 2014 annual conference on genetic and evolutionary computation. ACM
12. Khoshgoftaar TM et al (2013) A survey of stability analysis of feature subset selection techniques. In: 2013 IEEE 14th international conference on information reuse & integration (IRI). IEEE
13. Yang P et al (2013) Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. In: Biological knowledge discovery handbook: preprocessing, mining and postprocessing of biological data. Wiley, Hoboken, New Jersey, pp 333–52
14. 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}
15. Bolón-Canedo V, Alonso-Betanzos A (2019) Ensembles for feature selection: a review and future trends. Inf Fusion 52:1–12
16. Suriyamurthi D (2017) Stability of indexed microarray and text data. Int J Comput Algorithm 06(02):64–68
17. Xu Z et al (2017) An empirical study on the equivalence and stability of feature selection for noisy software defect data. SEKE
18. Alelyani S (2021) Stable bagging feature selection on medical data. J Big Data 8(1):1–18

19. Zhang L (2007) A Method for improving the stability of feature selection algorithm. In: Third international conference on natural computation (ICNC 2007), vol 1. IEEE
20. Kalousis A, Prados J, Hilario M (2005) Stability of feature selection algorithms. In IEEE international conference on data mining (ICDM'05)
21. Kamkar I, Gupta SK, Phung D, Venkatesh S (2015) Stable feature selection with support vector machines. In: Australasian joint conference on artificial intelligence (AI 2015), volume 9457 of LNCS, pages 298{308}
22. Ludwig L, Christoph M, Markus M, Kestler HA (2013) Measuring and visualizing the stability of biomarker selection techniques. Comput Stat 28(1):51{65}
23. Dhrif H et al (2019) A stable hybrid method for feature subset selection using particle swarm optimization with local search. In: Proceedings of the genetic and evolutionary computation conference. ACM
24. Park H et al (2019) Robust sample-specific stability selection with effective error control. J Comput Biol
25. Pes B (2019) Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput Appl 1–23
26. Tatwani S, Kumar E (2019) Effect of subset size on the stability of feature selection for gene expression data. World Congress on Engineering
27. Guidotti R, Ruggieri S (2018) Assessing the stability of interpretable models. arXiv preprint arXiv:1810.09352
28. Deraeve et al (2018) Fast, accurate, and stable feature selection using neural networks. Neuroinformatics 16(2):253–268
29. Arsov N, Pavlovski M, Kocarev L (2019) Stacking and stability. arXiv preprint arXiv:1901.09134
30. González J et al (2019) A new multi-objective wrapper method for feature selection—accuracy and stability analysis for BCI. Neurocomputing 333:407–418
31. Arsov et al (2019) Stability of decision trees and logistic regression. arXiv preprint arXiv:1903.00816
32. Cirillo M et al (2019) On the stability and generalization of learning with kernel activation functions. IEEE Trans Neural Netw Learn Syst arXiv preprint arXiv:1903.11990
33. Khaire et al (2019) Stability of feature selection algorithm: a review. J King Saud Univ Comput Inf Sci
34. Sechidis K et al (2019) On the stability of feature selection in the presence of feature correlations. In: European conference on "machine learning and principles and practice of knowledge discovery in databases" (ECML/PKDD)
35. Zhang et al (2018) A stability improved feature selection method for classification of ship radiated noise. OCEANS 2018 MTS/IEEE Charleston. IEEE
36. Suriyamurthi D, Velmurugan T (2018) Benchmarking attribute selection techniques for microarray data. ARPN J Eng Appl Sci 13:3740–3748
37. Goh WWB, Wong L (2016) Evaluating feature-selection stability in next generation proteomics. J Bioinform Comput Biol 14(05):1650029
38. Gaudelet T et al (2019) Unveiling new disease, pathway, and gene associations via multi-scale neural networks. arXiv preprint arXiv:1901.10005
39. 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):e28210
40. Chavez A et al (2019) Identify statistical similarities and differences between the deadliest cancer types through gene expression. arXiv preprint arXiv:1903.07847
41. 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
42. Bommert A, Lang M (2021) stabm: stability measures for feature selection. J Open Source Softw 6(59):3010
43. Zhang Y, Cao J (2020) GSimPy: a python package for measuring group similarity. SoftwareX 12:100526

44. Song X et al (2019) Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. J Am Med Inform Assoc 26(3):242–253
45. Zhang X, Jonassen, I (2018) EFSIS: ensemble feature selection integrating stability. arXiv preprint arXiv:1811.07939
46. Nogueira S (2018) Quantifying the stability of feature selection. The University of Manchester (United Kingdom), Diss