Journal of Big Data

RESEARCH Open Access

The stability of different aggregation techniques in ensemble feature selection

Reem Salman, Ayman Alzaatreh* and Hana Sulieman

*Correspondence:
aalzaatreh@aus.edu
Department of Mathematics and Statistics, American University of Sharjah, Sharjah 26666, UAE

Abstract

To mitigate the curse of dimensionality in high-dimensional datasets, feature selection has become a crucial step in most data mining applications. However, no feature selection method consistently delivers the best performance across different domains. For this reason and in order to improve the stability of the feature selection process, ensemble feature selection frameworks have become increasingly popular. While many have examined the construction of ensemble techniques under various considerations, little work has been done to shed light on the influence of the aggregation process on the stability of the ensemble feature selection. In contribution to this field, this work aims to explore the impact of some selected aggregation strategies on the ensemble's stability and accuracy. Using twelve classification real datasets from various domains, the stability and accuracy of five different aggregation techniques were examined under four standard filter feature selection methods. The experimental analysis revealed significant differences in both the stability and accuracy behavior of the ensemble under different aggregations, especially between score-based and rank-based aggregation strategies. Moreover, it was observed that the simpler score-based strategies based on the Arithmetic Mean or L2-norm aggregation appear to be efficient and compelling in most cases. Given the data structure or associated application domain, this work's findings can guide the construction of feature selection ensembles using the most efficient and suitable aggregation rules.

Keywords: Ensemble learning, Feature selection, Mean aggregation, Stability

Introduction

Through the development of information technology and the growing prevalence of data mining applications, machine learning has become a critical field of research and analysis over the last decade. Moreover, with the fast improvement of the efficient learning algorithms, the size of information that can be processed through machine learning has consistently evolved, often up to substantial proportions. Datasets that encompass large numbers of features often tend to be associated with higher levels of noise and larger potential for overfitting. The curse of dimensionality refers to a wide range of difficulties that arise from working with such high-dimensional data. In order to mitigate this issue, feature selection has become a necessary preprocessing step to the analysis of high-dimensional datasets [1]. Feature selection can be approached in a multitude of ways,

such as feature construction, feature ranking, multivariate feature selection, efficient search methods and feature validity assessment methods [2]. Based on how the method interacts with the learning algorithm, there are generally three classes of feature selection techniques . They are known as filter, wrapper and embedded methods [3]. Numerous reviews of the three categories can be found across the literature [4, 5].

The availability of a large number of feature selection methods has not yet brought about a general method that allows for extracting the perfect subset of features for building a model with optimum performance. Different feature selection algorithms identify different feature subsets for a given training data. Furthermore, a single feature selection algorithm can produce diverse top ranking feature subsets for varying data samples. In such cases, the feature selection process is termed *unstable*. The stability of the feature selection algorithms has received considerable attention in the recent data mining and machine learning literature [6, 7]. Stability or robustness in feature selection can be defined as the degree of variation in the selected feature subsets given small changes in the data used to obtain them. To enhance the stability of the selected feature subsets, *ensemble learning* has been developed and widely used in recent years.

Ensemble learning builds on the assumption that the aggregation of multiple models may provide better results than the use of a single model [8]. Ensembling in feature selection aggregates induced diversity by combining the results of multiple feature selection algorithms, generating and analyzing subsamples of the training data, or by combining both algorithm variation and data variation [9]. This can come at a larger computational cost, but may also provide better selected features and more stable results [10]. Although ensemble methods were originally introduced for classification models, Saeys et al. [10] proposed the idea of building ensemble frameworks for feature selection methods. Over the years, multiple works have highlighted the effectiveness of ensemble feature selection in producing efficient feature subsets, improving the performance of the learning algorithms, and elevating confidence in the obtained results and their practical implications [11–14]. Works such as [15, 16] have examined the effect of the ensemble frameworks on feature selection stability. However, most of these works tend to focus on the degree to which these ensembles perform in comparison to single selectors. To the best of our knowledge, not much focus has been given to the influence of the aggregation method itself on the stability and accuracy of the overall ensemble.

Thus, the objective of this paper is to study and single out the influence of the aggregation methods on both the classification performance and stability of the ensemble feature selection. To that end, this work employs the general bootstrap aggregation framework proposed recently by the authors in [9] which consists of generating multiple subsamples from the original dataset by bootstrapping, applying various feature selection techniques on each subsample and aggregating the results over algorithm variation and data variation. While the focus of this work is score-based aggregation for their continuous and detailed scale and the full distributional properties they possess, two rank-based aggregation strategies are included for completeness purposes and better insights of the comparison analysis conducted. Moreover, due to their computational efficiency, four traditional filter feature selection techniques are used in the experiments. The results of the extensive experimental work in this study demonstrate the efficiency of score-based aggregation methods such as the Arithmetic Mean and L2 Norm in

producing more stable or better performing feature subsets among other aggregation techniques. In addition, significant differences in the classification performance and stability behavior between the different aggregation methods have been deduced.

The rest of this paper is organized as follows. In "Background" section, we explore the relevant research background and compare with existing methods. In "Methodology" section, we briefly describe the bootstrap aggregation framework used in [9] and utilize it to introduce the framework for examining the stability influence of the aggregation techniques. In "Discussion" section, we present and analyze the results of the experimental work. Finally, conclusions and some insights for future work are presented in "Conclusion" section.

Background

Given a wide variety of feature selection methods, each with its own assumptions and rationale, most researchers agree that an optimal feature selection approach does not exist. Instead, they work on developing approaches that can handle large-scale datasets and provide optimal prediction accuracy for specific problems [17–19]. Among them, ensemble feature selection approaches are based on the notion that aggregating numerous models may yield better results than using a single model. Such ensemble frameworks typically encompass two main steps: First, a diversification approach is used to generate various feature selection outputs. The resulting outputs are then combined using an aggregation approach. The first step can be achieved through data variation, algorithm variation, or a hybrid of the two [20]. Data variation introduces perturbations in the dataset used, whereas algorithm variation achieves diversity by combining the outcomes of several feature selection methods. To optimize the efficacy of ensemble learning, ensemble feature selection frameworks have been studied under a variety of conditions. The second aggregation step, in particular, has a significant influence on the outcome of the ensemble feature selection [21].

Overall, several aggregation procedures have been employed across the literature. Most of the time, a rank aggregation technique is applied to combine obtained feature rankings from multiple feature selection methods [22–24]. Some of the most commonly used rank aggregation techniques include Borda's methods [25], Stuart [26], Robust Rank Aggregation [27] and SVM-Rank [28]. In their paper, Dittman et al. [29] applied nine rank aggregation techniques on an ensemble of twenty-five feature selection methods. Their findings highlight the effectiveness of rank aggregation, but indicate no significant differences between the different rank aggregation methods. A comparison of six rank aggregations techniques in [30] produced a similar conclusion. However, the minimum union method (Min), which simply selects the minimum of the obtained ranks, has proven more effective than others in some microarray datasets [31, 32]. In another study, Wald et al. [33] compared nine rank aggregation techniques across twenty-six bioinformatics datasets. An analysis of the results has led to clustering the different aggregation methods based on the similarity of their performance. Within the same cluster, the use of a simpler aggregation technique, such as mean rank aggregation in particular, could be recommended [34, 35].

While less commonly used in the literature, score-based aggregation is another approach for combining the output of multiple feature selection methods. In fact,

Borda's methods such as the Arithmetic Mean, Median, Geometric Mean and L2 Norm provide possible implementations. When information other than the ranks is available, Borda's score can be defined to take such additional information into consideration [25, 36]. Moreover, it was shown that applying the same Arithmetic Mean aggregation technique on both the feature importance scores and the feature ranks can lead to vastly different results [37]. Nevertheless, there is no clear rule on which of the two is the better option. In some literature, applying the same mean aggregation approach on both feature importance scores and feature rankings produced different outcomes depending on the data characteristics and application domains [38]. When compared using the relative weighted consistency as a stability measure, average score aggregation was found to be more stable than average rank aggregation. In addition, a positive association between stability and classification performance was discovered on artificial data [38]. In these cases, average score aggregation outperformed average rank aggregation. Alternatively, average rank aggregation outperformed average score aggregation when applied on the same datasets and compared using classification performance metrics in text categorization problems [39]. That is, as the aggregation strategy varies, so does the influence it can have on the ensemble construction in terms of both the classification performance and the stability.

Since numerous factors impact the stability of feature selection methods, including dataset variation [40], dataset imbalance [41], and feature redundancy [42]; several measures of feature selection stability have also been introduced [43–45], analyzed [46], or discussed across the literature [47]. These metrics can be used to determine more robust feature subsets for different datasets. In [48], Bommert and Rahnenführer proposed an adjusted stability measure by modifying the intersection scores between pairwise feature subsets, so that different but very similar features still contribute towards stability. As a result, this adjusted stability measure can be beneficial when working with datasets with highly correlated features. In [49], intensive search strategies such as the genetric algorithm were employed within ensembles to improve feature selection stability. However, while extensive published work concerns itself with comparing stability metrics and analyzing ensemble feature selection frameworks under different configurations, little work has been devoted to investigating the extent to which the aggregation techniques affect the stability of the ensemble feature selection on a general level. In [38], the relative weighted consistency is used to measure how frequently each feature is selected from among those that have been selected at least once and ignores the stability of non-selected features. The stability measure can then only be used for comparing subsets of selected features, rather than fully ranked lists. The purpose of this study is to fill some of the gap existing in the literature by investigating the influence of aggregating importance scores and feature rankings using various aggregation strategies on both the classification accuracy and stability of a bootstrap-based ensemble aggregation framework. The findings are meant to demonstrate the variations in aggregation strategies as well as the importance of the aggregation technique in influencing the overall ensemble feature selection. A merit of this work is that it examines both the potential differences in accuracy and stability between score-based and rank-based aggregations, and that these differences are assessed on multiple metrics.

Methodology

Based on diversified datasets generated from the original set of observations, Salman et al. [9] implemented a general ensemble framework in which the feature importance scores were generated by multiple feature selection techniques and aggregated using two methods: Within Aggregation Method (WAM) which refers to aggregating importance scores within a single feature selection and Between Aggregation Method (BAM) which refers to aggregating importance scores between multiple feature selection methods. This article uses the WAM algorithm in order to construct an ensemble framework and examine the influence of the aggregation method on both the stability and accuracy. In the next two subsections, we review the WAM approach used within the ensemble framework and describe the evaluation process for obtaining the feature selection stability. Then we construct a framework that incorporates the WAM algorithm within a cross-validation procedure in order to assess the influence of the aggregation strategy on the stability of the selected features.

Within aggregation method

Given a dataset $\mathbb{S} \equiv (\boldsymbol{X}, Y)$, with n observations and p features, such that $\boldsymbol{X} = [x_{ij}]_{n\times p} \in \mathbb{R}^{n,p}$ is the data matrix and Y is the target variable. Moreover, let $\{V_1, \ldots, V_p\}$ denote the set of features (variables) in $\boldsymbol{X}$. The entire dataset $\mathbb{S}$ is then divided into a training dataset $\mathbb{X}$ and a testing dataset $\mathbb{T}$. First, multiple training subsamples are generated by bootstrapping $\mathbb{X}$. Subsequently, a number of feature selection techniques $FS_1, \ldots, FS_t$ are applied on each subsample, generating a feature importance score $\ell_j \in \mathbb{R}$ for every feature $V_j \in \{V_1, \ldots, V_p\}$. The Within Aggregation Method (WAM) is used for aggregating the *importance scores* within a single feature selection method, for each of the feature selection methods used. Based on the aggregated importance scores, the feature set is then sorted from the most to the least important to obtain a rank vector $\boldsymbol{r} = (r_1, \ldots, r_p)$, $r_j \in \{1, 2, \ldots, p\}$. For a given threshold parameter value k, where $0 < k \leq 1$, then only the most important $100k\%$ of the feature set is retained and used to test the model performance using a learning algorithm. Different feature selection methods can be utilized and compared to choose the best feature selection approach for a given problem. The WAM procedure is illustrated in Figure 1 and Algorithm 1 below details its implementation.

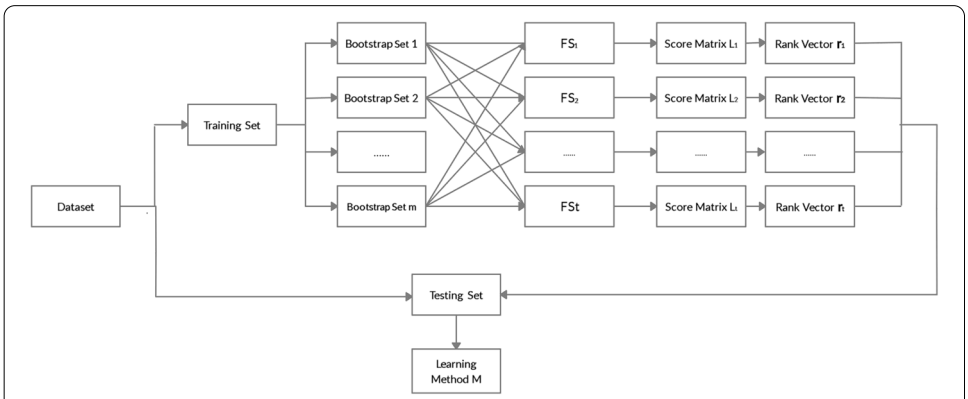

Fig. 1 Framework for WAM

Algorithm 1 WAM Algorithm

Given a training dataset $\mathbb{X}$ with p features, a testing dataset $\mathbb{T}$, a feature selection method FS, a threshold parameter k, and a learning algorithm M.

1: For $s = 1, \ldots, m$, generate bootstrap samples, $\mathbb{X}_1, \ldots, \mathbb{X}_m$ of the training dataset $\mathbb{X}$.
2: Based on FS, get the features score matrix $\mathbb{L}$.
3: Get the aggregated score set $\{a_1, \ldots, a_p\}$.
4: For the aggregated score set $\{a_1, \ldots, a_p\}$, get the corresponding rank vector $\boldsymbol{r} = (r_1, \ldots, r_p)$.
5: Based on the rank vector $\boldsymbol{r}$, keep only the top $100k\%$ of the variable set $\{V_1, \ldots, V_p\}$.
6: Based on the selected feature set in (5), use the testing dataset $\mathbb{T}$ and a cross-validation technique to train and test the model M.

Feature selection stability

The stability of feature selection generally refers to the robustness of the obtained feature subsets against variations in the training data. Unstable feature selection methods are significantly sensitive to such data variations and thus produce different feature subsets across training sets derived from the same dataset. This raises doubts about the reliability of the feature selection even if it is accompanied with optimal learning performance. For this reason, understanding and studying the stability of feature selection has become crucial in many areas of application. In this section, we briefly describe some of the principal measures used to assess the stability of a feature selection method. In the following "Stability influence of the aggregation technique" section, we adapt such stability measures to develop the scheme for investigating the stability influence of the aggregation techniques within the bootstrap aggregation framework [9].

Since the stability of feature selection is mainly affected by data variation, there are two common approaches for inducing variation in the data [47]:

> **Approach 1** Divide the dataset $\mathbb{S}$ into a training dataset $\mathbb{X}$ and a testing dataset $\mathbb{T}$. Then, multiple training samples $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_m$ are generated by bootstrapping $\mathbb{X}$ as described in the WAM framework (Algorithm 1).
>
> **Approach 2** Use m-fold cross-validation to generate different training datasets. For example, take a 5-fold cross-validation procedure. On every iteration, one of these folds is used as a testing dataset $\mathbb{T}$, while the remaining four folds are used as a training dataset $\mathbb{X}$. In this manner, by going through all iterations, one can obtain five training samples $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_5$ and five testing samples $\mathbb{T}_1, \mathbb{T}_2, \ldots, \mathbb{T}_5$.

For both approaches, evaluating the stability of a feature selection method can be done by simply taking the average of similarity comparisons between every pair of feature rankings derived from different training samples as follows:

$$Stability = \frac{2}{m(m-1)} \sum_{s=1}^{m-1} \sum_{v=s+1}^{m} \Phi(f_s, f_v) \tag{1}$$

where m is the total number of training samples and $\Phi(f_s, f_v)$ is the similarity measure between a pair of feature rankings from any two training samples $\mathbb{X}_s, \mathbb{X}_v$ $(1 \leqslant s, v \leqslant m)$. Note that in order to compute the similarity measure $\Phi(f_s, f_v)$, the pair (f_s, f_v) can be

represented as two importance scores vectors, two rank vectors, or two index vectors. Moreover, several similarity measures have been introduced across the literature for each of the aftermentioned representations [50]. In this paper, we obtain rank vectors $\{\boldsymbol{r_1}, \ldots \boldsymbol{r_m}\}$ from our feature selection methods and use them to obtain the following three similarity measures:

i. *Spearman Rank Correlation Coefficient*. Spearman's Rank Correlation Coefficient measures the similarity between the two rank vectors as:

$$\Phi_{SRCC}(\boldsymbol{r_s}, \boldsymbol{r_v}) = 1 - \frac{6\sum_{j=1}^{p}(r_{sj} - r_{vj})^2}{p(p^2-1)}, \tag{2}$$

where $\boldsymbol{r_s}$ is the rank vector that corresponds to the set of features $\{V_1, \ldots, V_p\}$, such that $\boldsymbol{r_s}$ is derived from the importance scores $\boldsymbol{\ell_s}$. Here, $\Phi_{SRCC}(\boldsymbol{r_s}, \boldsymbol{r_v}) \in [-1, 1]$.

ii. *Canberra Distance*. Cenberra's Distance represents the absolute difference between two rank vectors as:

$$\Phi_{CD}(\boldsymbol{r_s}, \boldsymbol{r_v}) = \sum_{j=1}^{p} \frac{|r_{sj} - r_{vj}|}{|r_{sj}| + |r_{vj}|}. \tag{3}$$

where $\boldsymbol{r_s}$ is the rank vector that corresponds to the set of features $\{V_1, \ldots, V_p\}$, such that $\boldsymbol{r_s}$ is derived from $\boldsymbol{\ell_s}$. For easier interpretation, Canberra's distance is normalized through dividing by p.

iii. *Jaccard's index*. Jaccard's index measures the similarity between two finite sets; it is taken as the size of the intersection divided by the size of the union of the two sets as follows:

$$\Phi_{JI}(\boldsymbol{w_s}, \boldsymbol{w_v}) = \frac{|\boldsymbol{w_s} \cap \boldsymbol{w_v}|}{|\boldsymbol{w_s} \cup \boldsymbol{w_v}|} = \frac{|\boldsymbol{w_s} \cap \boldsymbol{w_v}|}{|\boldsymbol{w_s}| + |\boldsymbol{w_v}| - |\boldsymbol{w_s} \cap \boldsymbol{w_v}|}, \tag{4}$$

where $(\boldsymbol{w_s}, \boldsymbol{w_v})$ are index vectors which are used to represent a pair of feature subsets. By selecting the top $100k\%$ features, any two rank vectors $(\boldsymbol{r_s}, \boldsymbol{r_v})$ can be converted into the index vectors $(\boldsymbol{w_s}, \boldsymbol{w_v})$. Here, $\Phi_{JI}(\boldsymbol{w_s}, \boldsymbol{w_v}) \in [0, 1]$.

Stability influence of the aggregation technique

To assess the stability influence of the aggregation technique on the ensemble framework, we combine the two approaches, Approach 1 and Approach 2, mentioned in "Feature selection stability" section. As shown in Figure 2, we embed the WAM framework, constructed using Approach 1, inside a cross-validation procedure of Approach 2. That is, at every iteration of the cross-validation procedure, there exists an internal WAM framework in which an aggregated feature selection result is obtained. To do this, we first take the dataset $\mathbb{S}$ and apply cross-validation to generate f-folds of the dataset. On every iteration, 1 fold is taken aside (dataset $\mathbb{T}$) whereas the remaining $f-1$ folds are used for training (dataset $\mathbb{X}$). We then apply Algorithm 1 (WAM) on the training data. This means that the bootstrap samples $\mathbb{X}_1, \ldots, \mathbb{X}_m$ will be generated inside every iteration of the cross-validation and used to obtain an aggregated rank vector $\boldsymbol{r}$. By going through all iterations, we obtain $\{\boldsymbol{r_1}, \ldots \boldsymbol{r_f}\}$ aggregated rank vectors.

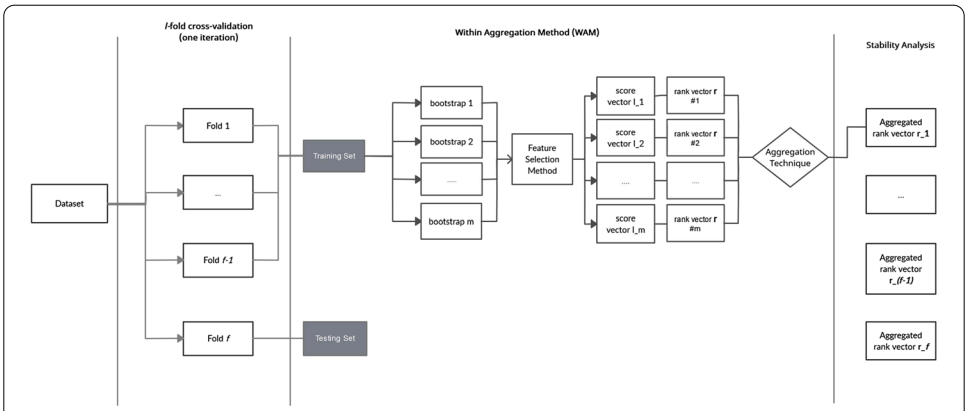

Fig. 2 Framework for Stability Analysis

Given the aggregated rank vectors, $\{\boldsymbol{r_1}, \ldots \boldsymbol{r_f}\}$, the stability of the ensemble feature selection can then be computed by averaging over the values of any of the similarity measures described in "Feature selection stability" section, i.e., average Spearman's Rank Correlation, average Canberra's Distance and average Jaccard's Index. That is, based on a selected similarity measure, the similarity scores of all the pairings of the aggregated feature rankings are computed and then averaged to obtain the final stability score. Since the WAM is embedded within the cross-validation procedure, each output of the WAM produces an aggregated rank vector for each aggregation technique. Data variation is introduced to the these aggregated rank vectors through the cross-validation folds (i.e. Approach 2). Because a single feature selection method is implemented across all the folds, the resulting volatility is dominantly attributed to the aggregation technique within the ensemble; hence, controlling for the effect of the aggregation strategy on the stability and robustness of the ensemble feature selection.

Experimental evaluation

In this section, multiple experiments are performed using several feature selection algorithms and various score-based aggregation techniques. For the sake of completeness of the discussion and comparisons, we have added two rank-based aggregation techniques to the experimental work. Also, to allow for better generalization, this paper implements the described methodology on twelve real datasets from different domains. While the datasets used fall under classification problems, the framework is also applicable for regression problems. The following two subsections provide further details of the experimental work and evaluation.

Experimental datasets

For each of the considered datasets, Table 1 reports a brief description of the overall number of observations, features and classes in the target variable Y. As can be noticed, the experimental framework consists of binary and multiclass datasets. Moreover, the dimensionality and class balance distributions varied across the tested datasets. The number of features ranges from 34 to 309 and the number of observations ranges from 351 to 9298 observations. In addition, all of them are real data which can be downloaded

Table 1 Datasets description

Dataset name and source	No. observations	No. Features	No. Classes	Dimensionality*
Jasmine[a]	2984 (1492/1492)	145	2	0.048592
Image[b]	2000 (1420/580)	140	2	0.070000
Scene[c]	2407 (1976/431)	295	2	0.122559
Musk[d]	6598 (5581/1017)	170	2	0.025765
Philippine[a]	5832 (2916/2916)	309	2	0.052984
Ionosphere[d]	351 (126/225)	34	2	0.096866
Optdigits[b]	5620 (572/5048)	64	2	0.011388
Satellite[b]	5100 (75/5025)	37	2	0.007255
Ada[a]	4147 (1029/3118)	49	2	0.011816
Splice[b]	3190 (1535/1655)	62	2	0.019436
Indian Pines[b]	9144	221	8	0.024168
Semeion[d]	1593	257	10	0.161330

*Dimensionality is the ratio of features to number of observations. Superscripts indicate the data sources as follows:

[a] automl.chalearn.org

[b] www.openml.org

[c] mulan.sourceforge.net

[d] archive.ics.uci.edu

from the provided links. These datasets come from different application domains and provide a useful benchmark for the experimental evaluation.

Experimental design

In this experiment, we first study the classification performance of different aggregations strategies under the Within Aggregation Method (WAM). To do this, the following four filter feature selection methods are implemented:

- Information Gain (IG): Due to its computational efficiency and simple interpretation, information gain is one of the most popular feature selection methods [51]. It is a symmetrical measure of dependency between two random variables X and Y, which quantifies the information gained about Y after observing X, or vice versa. In supervised learning, Y is taken to be the target variable. Thus, the aim of IG is to measure how much information a feature gives about the target variable.
- Symmetric Uncertainty (SU): A correlation measure between a pair of random variables, usually an attribute in a dataset X and the target variable Y. Similar to Information Gain, SU depends on entropy to gauge a variable's information. However, the SU criterion compensates for the inherent bias of Information Gain by dividing it by the sum of the entropies of X and Y.
- Chi-Square (CH): The Chi-Square test is implemented as a way of testing the independence of two discrete variables, by examining whether the observed distributions are generated by the same underlying distribution. This statistic depends on the difference between the observed and expected class frequencies, the degree of freedom and the samples size. To implement this technique, numeric features were discretized based on a fixed-width binning.

- Minimum Redundancy Maximum Relevance (MRMR): The aim of this method is to select features that are highly correlated with the target variable (maximum relevance), yet show little correlation between the features themselves (minimum redundancy).

In each experiment, two-thirds of the dataset were used for the training phase and one third for the testing. Moreover, 1000 bootstrap samples were generated to aggregate the feature importance scores within each feature selection method. Accordingly, the five following aggregation techniques are utilized: Arithmetic Mean, Geometric Mean, L2 Norm, Robust Rank Aggregation (RRA) and Stuart aggregation as illustrated in Table 2. It should be noted here that the first three aggregation techniques combine the importance scores obtained from the feature selection, whereas in RRA and Stuart, the importance scores are converted into rank vectors first and then the ranks are combined. In comparison to importance scores, rank-based aggregations are more resistant to outliers and invariant to transformation and normalization. Score-based aggregations, on the other hand, restrict the information loss required in generating the ranks and can be easier to obtain. In addition, because aggregated score vectors can be readily turned into rank vectors and feature subsets, they may be used with a larger range of stability metrics.

To determine the classification accuracy of the data, a five-fold cross-validation procedure was implemented in the testing. The learning algorithm utilized was Naive Bayes, a probabilistic classifier which assumes the occurrence of each input feature is independent from others and is applicable for both binary and multiclass problems. For the testing stage, 10 different k thresholds were used, resulting in subsets containing the top {10%, 20%,..., 100%} of the total features.

Next, we compared the influence of the five different aggregation techniques on the stability of the overall ensemble feature selection. In the experiments, a 100-fold cross-validation procedure was implemented. On every iteration, 99 folds were used to obtain a training set $\mathbb{X}$ and one fold was used to obtain a testing set $\mathbb{T}$. Each training set underwent a WAM bootstrap aggregation framework, to obtain final aggregated

Table 2 Details of aggregation techniques

Aggregation Technique	Formula	Description
Arithmetic Mean	$\frac{a_1 + a_2 + \ldots + a_m}{m}$	Calculates the average across importance scores and uses it to determine final aggregated score [36]
Geometric Mean	$\sqrt[m]{a_1 a_2 \ldots a_m}$	Calculates the geometric average across importance scores and uses it to determine final aggregated score [36]
L2 Norm	$\sqrt{a_1^2 + a_2^2 + \ldots + a_m^2}$	Views the importance scores as an n-dimensional vector and calculates the Euclidean norm for that vector [36]
Stuart	$\Pr[X \leq \rho] = 1 - \Pr[\hat{r}_{(1)} \leq 1 - \beta^{-1}_{m,m}(\rho), \ldots, \hat{r}_{(m)} \leq 1 - \beta^{-1}_{m,1}(\rho)]$	Compares obtained rank vectors to a baseline of randomly ranked features then assigns the features significance scores using the beta distribution [26]
RRA	$\rho(r) = \min_{k-1} \beta_{k,m}(r)$, $\beta_{k,m}(x) := \sum_{\ell=k}^{m} \binom{\ell}{m} x^{\ell}(1-x)^{m-\ell}$	Similar to Stuart, but achieves efficiency & precision trade-off by using Bonferroni corrections [27]

rank vectors using the different aggregation techniques. In other words, 100 iterations produced 100 aggregated rank vectors for each of the five aggregation techniques. These final 100 rank aggregated rank vectors were then used to compute the stability of the ensemble feature selection under each aggregation method. This process was also applied for each of the two feature selection methods: IG and MRMR. These two methods were selected due to exhibiting the most contrasting behavior in the current classification accuracy evaluation and in previous work [9].

The entire experimental framework was performed on the open-source statistical programming language R. The experimental environment in which the testing took place was Windows 10, 64-bit, 16 GB RAM, Intel(R) Xeon E-2124 (3.30GHz). Note that features of near-zero variance were removed prior to the analysis.

Discussion

Figures 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 illustrate the resulting accuracy (proportion of correct predictions) after applying the WAM on each dataset using different aggregation techniques. Each of the curves corresponds to the aggregation methods: Arithmetic Mean, Geometric Mean, L2 Norm, Robust Rank Aggregation (RRA) and Stuart aggregation. On the other hand, the four graphs for each dataset depict the four feature selection methods used: Information Gain, Symmetric Uncertainty, Chi-Squared and MRMR. The classification accuracy values are averaged over the 5-folds in the cross-validation and plotted against the ten different 100k% thresholds. In Figs. 15, 16, 17, 18, 19, 20, the stability score resulted from each of the aggregation methods is depicted against the 12 datasets, using the three stability metrics: Average Jaccard's Index, average Spearman's Rank Correlation and average Canberra's Distance (1-average Canberra's Distance is used in Figs. 15, 16, 17, 18, 19, 20 so that higher values on the y-axis indicate higher stability in accordance with the other two stability metrics).

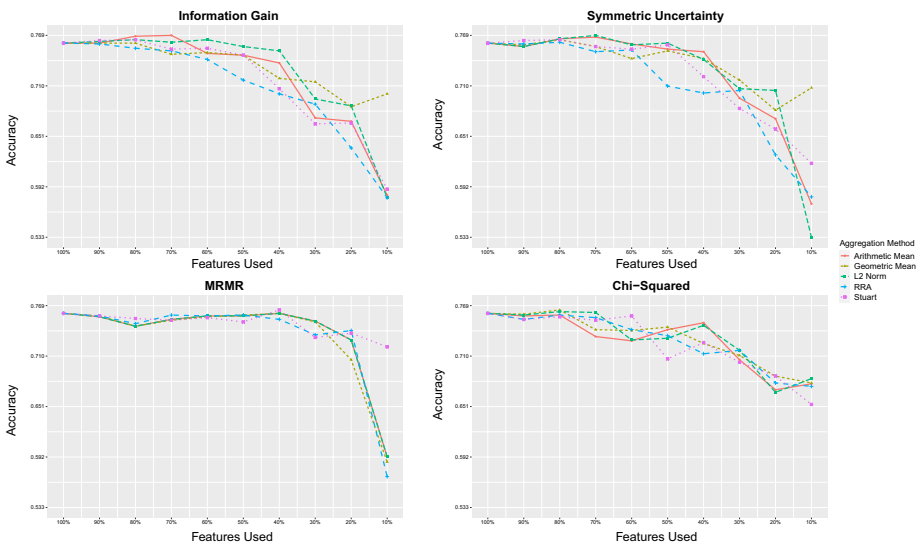

Fig. 3 Jasmine dataset classification results

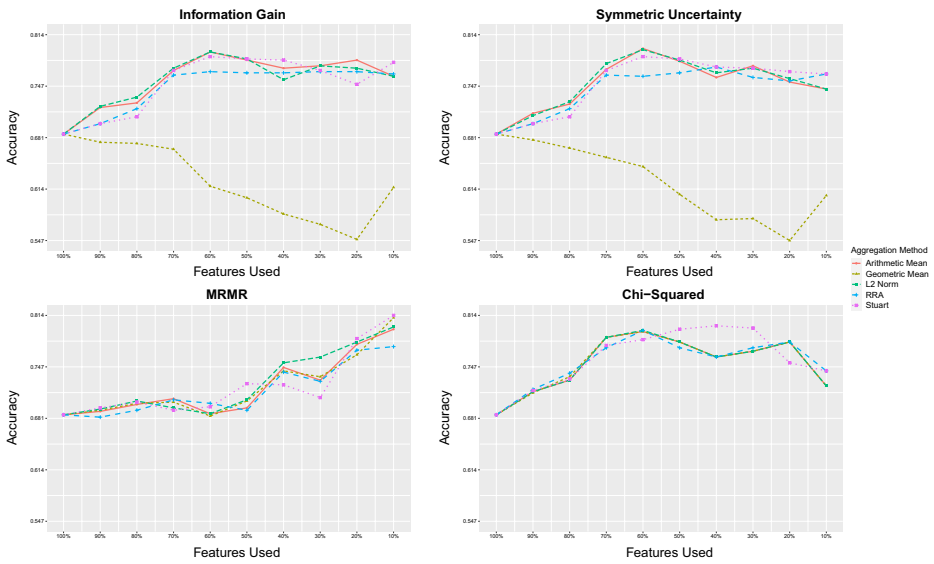

Fig. 4 Image dataset classification results

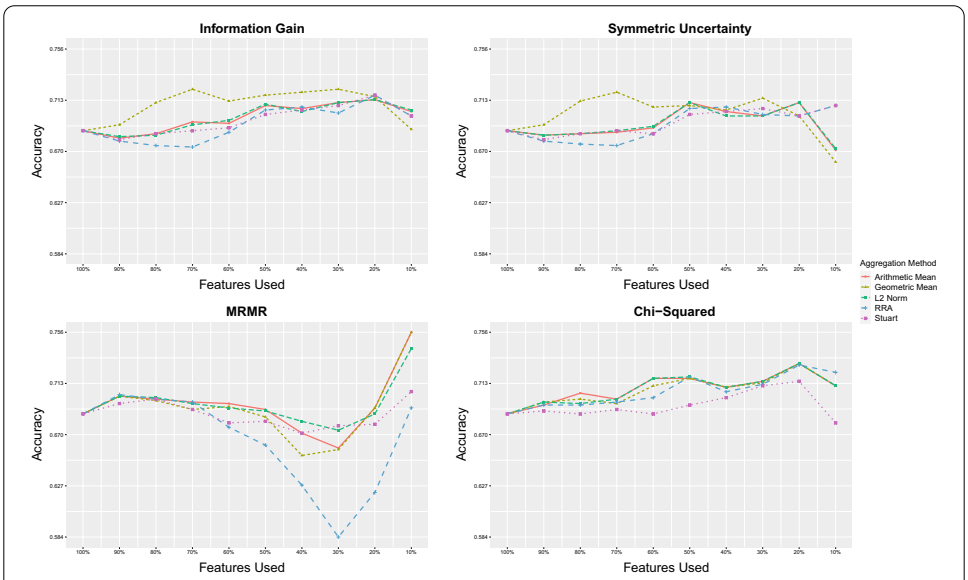

Fig. 5 Scene dataset classification results

The three stability metrics were calculated for the ensembles constructed using both Information Gain and MRMR and depicted by three separate plots for each feature selection method. In the next two subsections, we analyze and compare the classification performance and stability influence of the five aggregation methods.

Classification performance (Figures 3-14)

It is clearly observed that the WAM improves the baseline (100%) classification performance under most aggregations. That is, the accuracy increases in almost each of

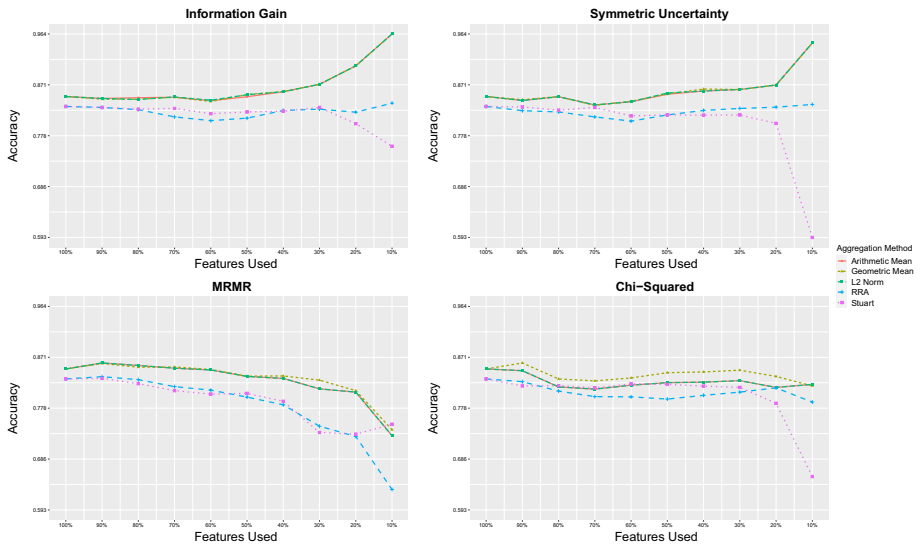

Fig. 6 Musk dataset classification results

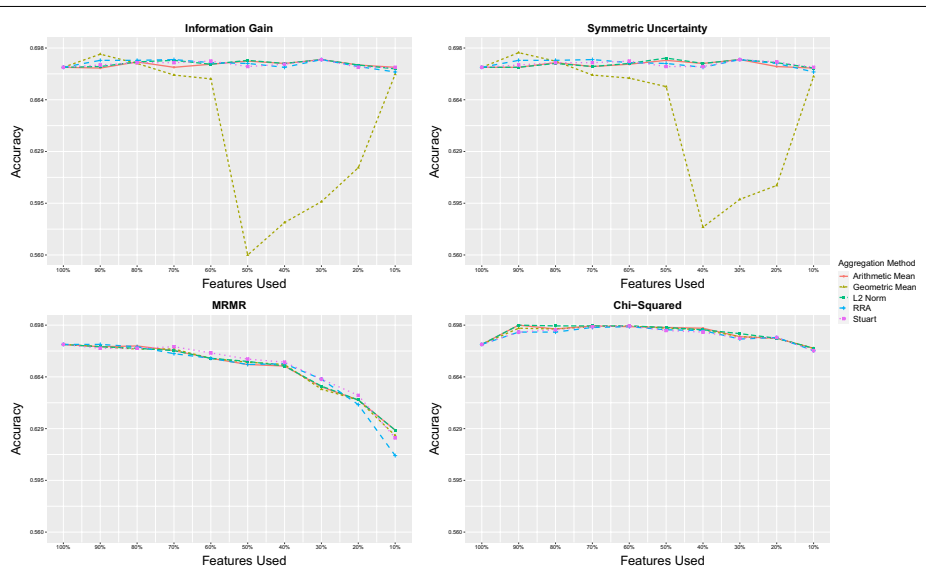

Fig. 7 Philippine dataset classification results

Figures 3-14 after selecting the most relevant features using one of the implemented feature selection methods. Moreover, we note that the curves overlap for most aggregations under higher feature selection thresholds. This is consistent with the findings in [33], in which it was observed that the similarity between the aggregation methods effectively increases as the selected feature subsets grow larger. Since the number of common features between the aggregated rank vectors will increase as more features are considered, the obtained results are unsurprising. The general trend also illustrates a sharp decrease in the achieved accuracy behavior once most of the dataset features are removed. In particular, this is often pronounced when at most 30% (e.g. Ionosphere, Ada, Splice) or 20% (e.g. Jasmine, Spectrometer, Musk) of the total features are retained. Accordingly, the

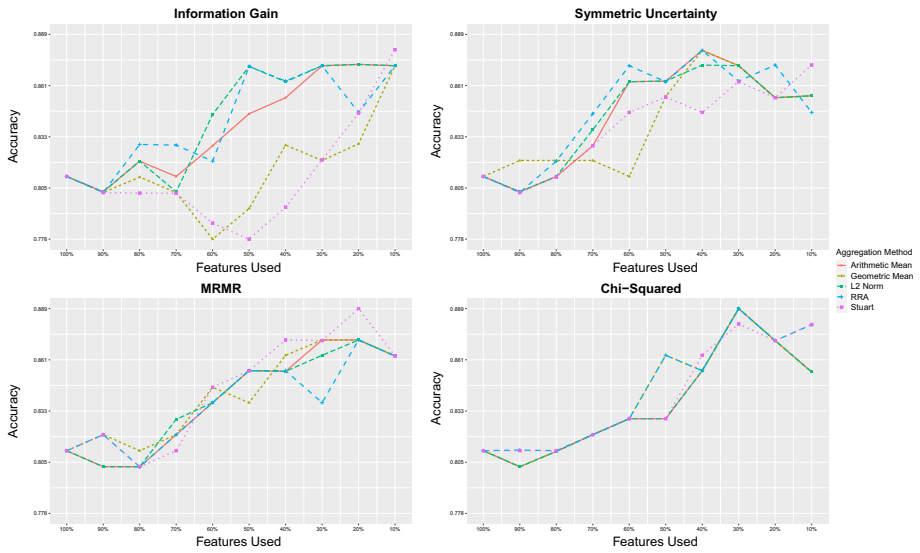

Fig. 8 Ionosphere dataset classification results

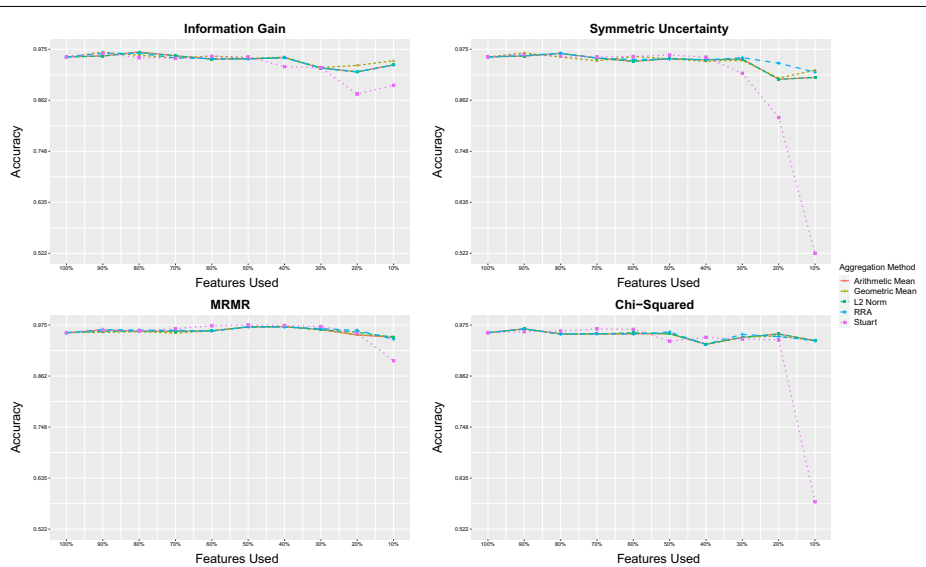

Fig. 9 Optdigits dataset classification results

pattern still depends on the data itself, with the optimal feature reduction threshold varying across the experiments. These findings promote the utilization of the WAM framework in order to find the most suitable feature selection method or guide the selection of an optimal feature reduction threshold.

Furthermore, the classification performance under each of the aggregation methods can be viewed as data-dependent. For example, while Geometric Mean aggregation does well in the Scene and Musk datasets, it is one of the least performing aggregation frameworks under Image, Ionosphere and Philippine datasets. A similar pattern can be seen with Stuart rank aggregation. These contrasts are also characterized by the nature of the feature selection methods tested. A repeated-measures ANOVA reveals significant

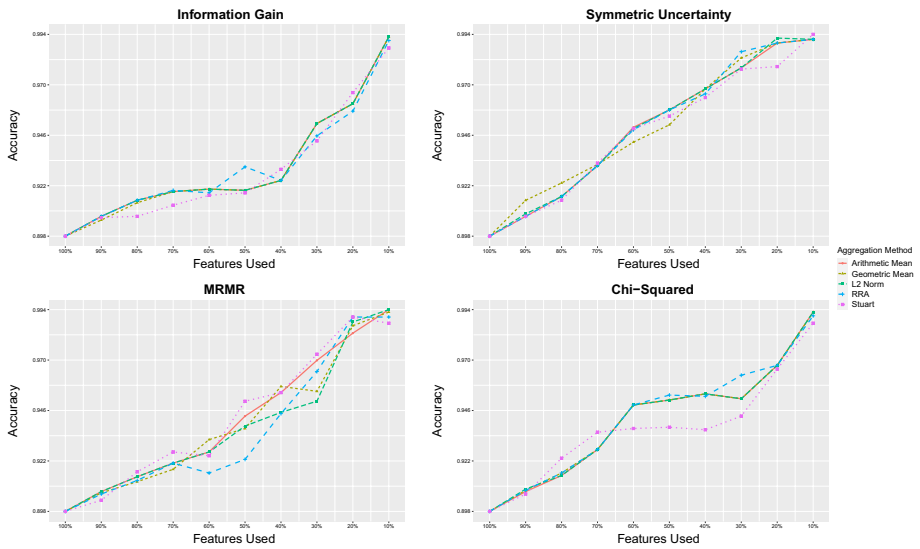

Fig. 10 Satellite dataset classification results

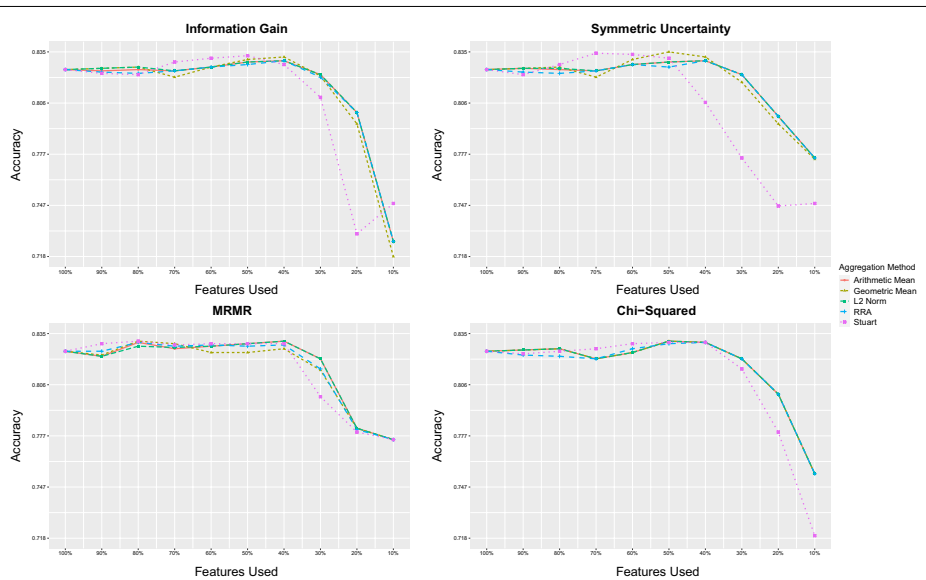

Fig. 11 Ada dataset classification results

differences in the aggregation accuracies under both Information Gain and MRMR methods (Bonferroni-adjusted p-values 0.0068 and 0.0304 respectively). When further investigated, the significant pairwise differences were generally attributed to a difference between one of the score-based aggregations and the rank-based aggregations. For example, under MRMR, significant accuracy differences were reported between the rank-based RRA and all of the score-based aggregations Arithmetic Mean, Geometric Mean and L2 Norm (adjusted p-values 0.001, 0.005 and 0.002 respectively). Otherwise, significant pairwise differences were also observed between the Geometric Mean and

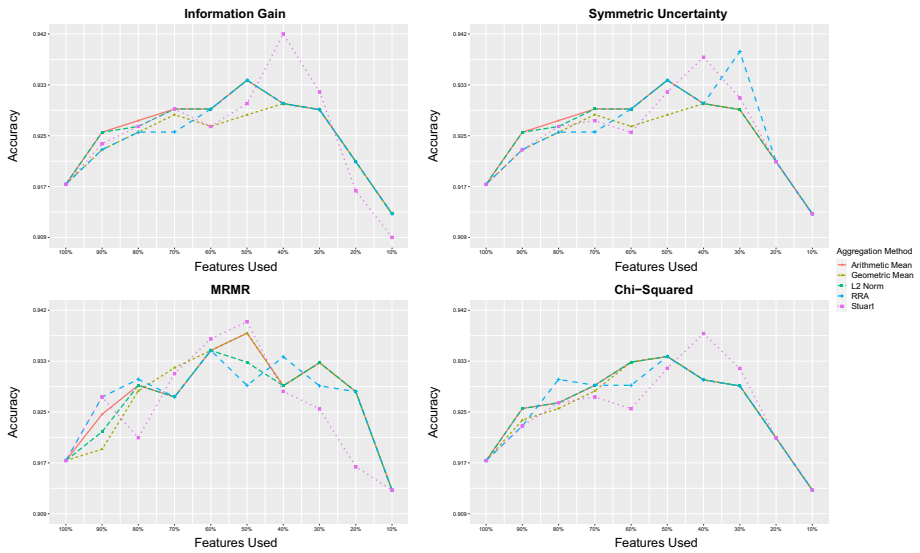

Fig. 12 Splice dataset classification results

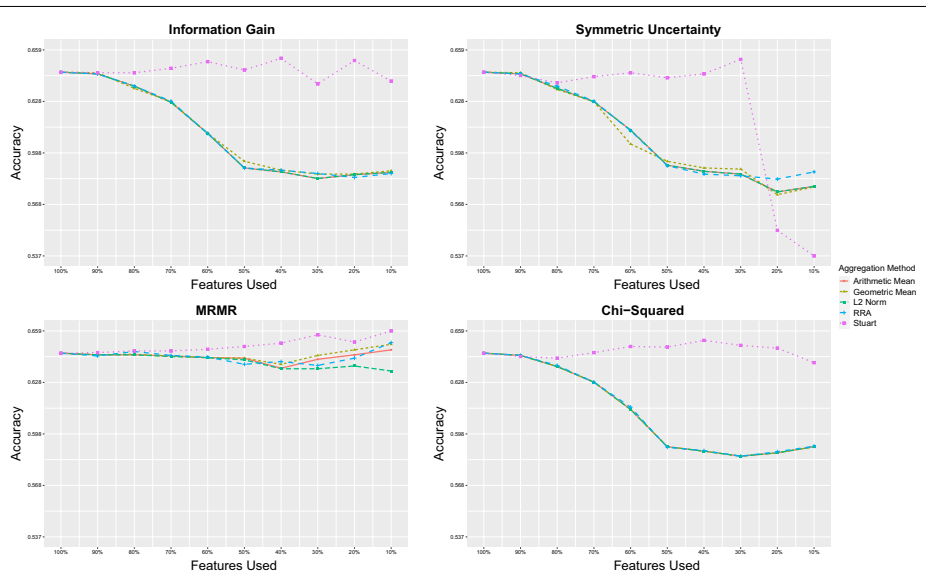

Fig. 13 Indian Pines dataset classification results

Arithmetic Mean. This does come off as unexpected, since the Geometric Mean presents a rather volatile performance across the datasets (e.g. Image, Scene, Philippine).

In contrast, we note that both the score-based Arithmetic Mean and L2 Norm aggregations behave noticeably well across most examples. In 11 out of 12 datasets, both aggregation techniques are consistently the middlemost or the topmost performing in comparison to the other aggregations. Furthermore, both methods exhibit much steadier accuracy performance across all the different datasets. When compared with each other, we observe that the Arithmetic Mean is slightly more consistent than the L2 Norm especially in Ionosphere and Satellite datasets, though the curves frequently overlap in terms of classification accuracy. Given their favourable

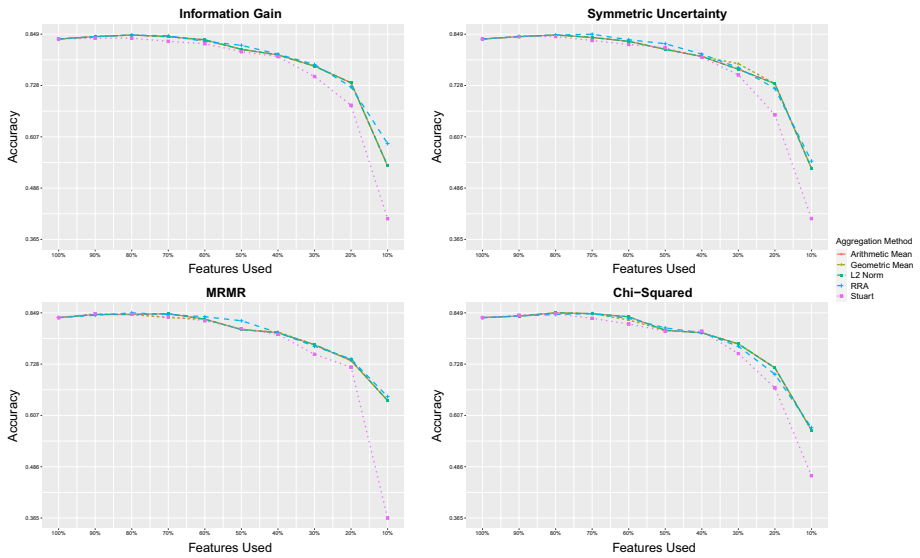

Fig. 14 Semeion dataset classification results

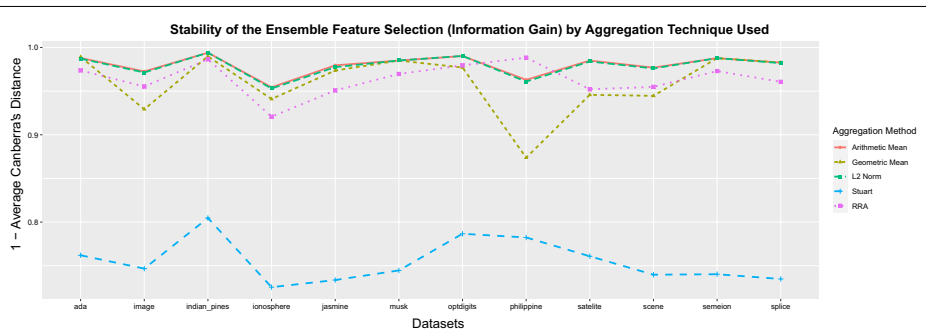

Fig. 15 Stability Analysis using Canberra's Distance (IG)

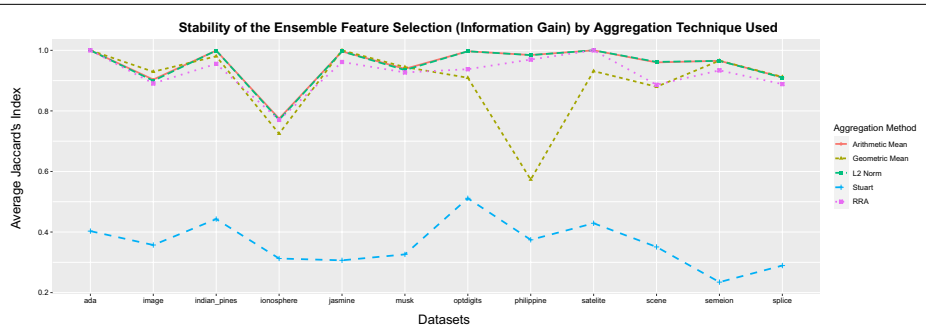

Fig. 16 Stability Analysis using Jaccard's Index (IG)

behavior in these experiments, either of the two aggregations can be utilized as simple and efficient techniques for achieving effective classification results under the WAM framework. In particular, the Arithmetic Mean is a reliable choice considering its simplicity and ease of implementation. These findings are consistent with previous

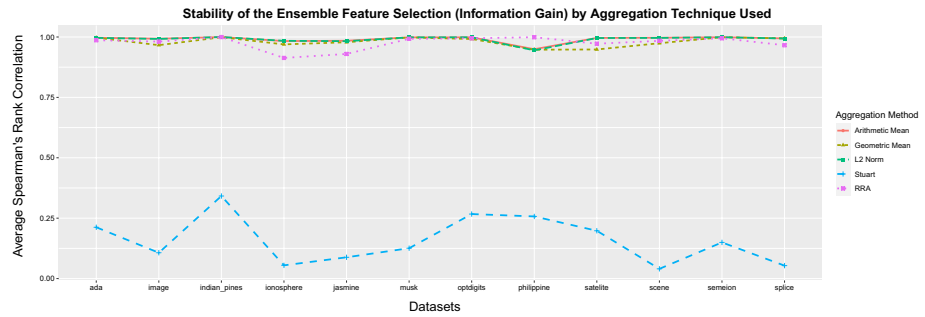

Fig. 17 Stability Analysis using Spearman's Rank Correlation Coefficient (IG)

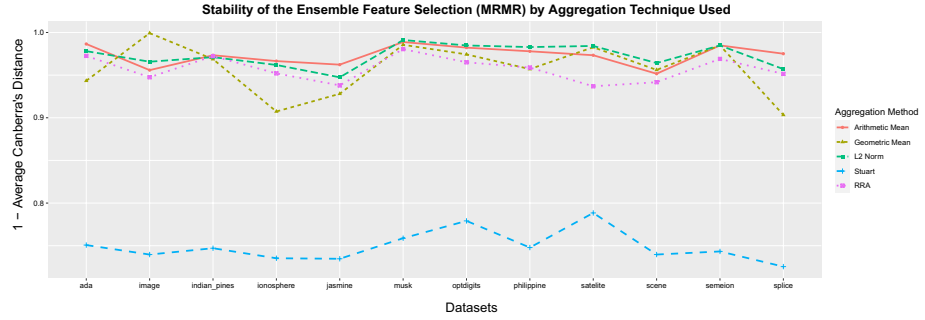

Fig. 18 Stability Analysis using Canberra's Distance (MRMR)

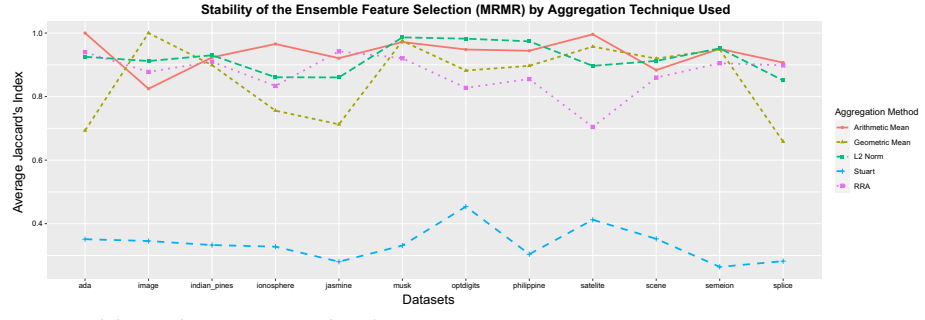

Fig. 19 Stability Analysis using Jaccard's Index (MRMR)

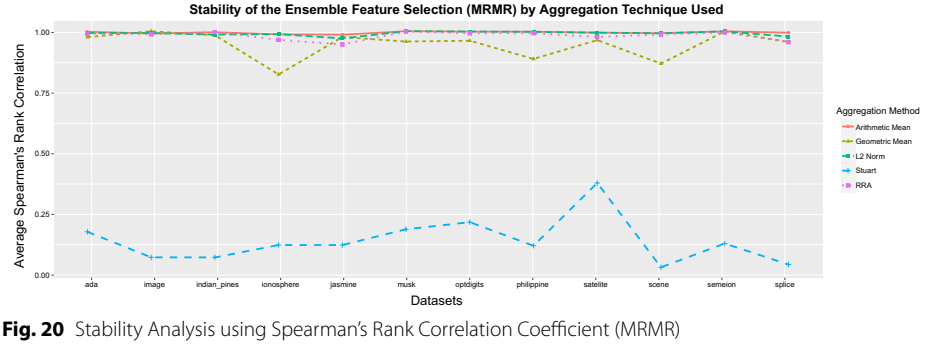

Fig. 20 Stability Analysis using Spearman's Rank Correlation Coefficient (MRMR)

works that demonstrate the efficiency of the Arithmetic Mean as a general aggregation rule [9, 34, 35].

In summary, the examination of the classification accuracy results reveals that there exist significant differences between the aggregation techniques, particularly the score-based and rank-based aggregations. Also a significant difference is seen between the Arithmetic Mean and the Geometric Mean aggregations while the difference between the Arithmetic Mean and L2 Norm is insignificant. In fact, for most of the experiments, the Arithmetic Mean and L2 Norm appear to outperform other aggregation rules in both accuracy and robustness over different feature selection methods.

Stability analysis (Figures 15-20)

Interestingly, the stability behaviour highlighted by the five aggregation methods depicted in Figures 15-20 has a steadier pattern than that observed in the classification performance. Under both Information Gain and MRMR, we notice that Arithmetic Mean and L2 Norm generally produce the highest stability scores, peaking at multiple datasets (e.g. Jasmine, Optdigits, Indian Pines) in comparison to the other aggregation methods. Moreover, the rank-based Stuart aggregation produces the lowest stability scores across all the experimental datasets and for all the implemented stability metrics. Overall, the two feature selection methods, Information Gain and MRMR, exhibit nearly similar behavior in their stability performance. For the stability metrics themselves, Jaccard's Index demonstrates the least consistent pattern over the five aggregation methods across all the datasets. However, there was little difference between the binary and multiclass dataset results.

Statistically, one-way ANOVA indicates that there exist significant differences in the mean stability scores represented by the different aggregations in each of the six figures (Bonferroni-adjusted p-values < 0.0001). Post-hoc investigation reveals that, in accordance with the graphs, the significant differences can almost always be attributed to Stuart aggregation in comparison to any other aggregation method (Bonferroni-adjusted p-values<0.0001). Other significant differences are also found under Information Gain using Canberra's Distance between the Arithmetic Mean and L2 Norm (Bonferroni-adjusted p-value$=0.037$), Arithmetic Mean and RRA (Bonferroni-adjusted p-value$=0.041$), and using Jaccard's Index between the Arithmetic Mean and RRA (Bonferroni-adjusted p-value$=0.031$). As in the classification accuracy performance, the Arithmetic Mean and L2 Norm do not differ as much in their influence over the stability of the ensemble. Overall, in these experiments, the Arithmetic Mean and L2 Norm appear to be the better choice in terms of stability compared to more complex alternatives. As the scores have a stronger scale and provide a higher level of details about the importance of the features, the findings of this analysis highlight the importance of recognizing the differences in stability influence between score-based and rank-based aggregations in the construction of the ensemble feature selection framework.

It is worth noting that for any vector of nonnegative real numbers $a_1, a_2, \ldots, a_n$, it can be shown that the three score-based aggregation techniques used in this paper exhibit the following relationship:

$$\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2} \geq \frac{a_1 + a_2 + \ldots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \ldots a_n}$$

In other words, the Arithmetic Mean of any vector lies between the L2 Norm and its Geometric Mean [52]. However, although these score-based aggregations can be computed in parallel, it is still possible for each of the aggregations to produce different rank vectors due to different sorting of the aggregated importance scores themselves. With this understanding, some of the score-based aggregations within the experimental analysis in this work reveal behavioral dissimilarities. The above inequality showing the relationship between the three aggregation techniques can actually provide some understanding of the extent in which the inherent variability of each technique can influnece the stability of the feature selection process.

Conclusion

With the growing prevalence of big data applications, feature selection has become a necessary preprocessing tool across many domains. Ensemble feature selection has emerged as a new data mining method with the premise of improving both feature selection stability and learning algorithm performance. Several measures of feature selection stability in ensemble learning have been recently introduced and analyzed in the literature. However, little work has been done to investigate the stability behavior of different aggregations within the same ensemble feature selection. In this work, we have investigated the behavior of the ensemble feature selection in terms of learning accuracy and stability under variation in the aggregation process based on repeated bootstrap sample generation. The effect of five different aggregation techniques was experimentally evaluated under four traditional filter feature selection methods using twelve classification real datasets from various domains. The five techniques included three score-based aggregation: Arithmetic Mean, Geometric Mean and L2 Norm, and two rank-based aggregation: RRA and Stuart.

The findings of the experimental evaluation highlighted the merits of the simple score-based aggregations such as the Arithmetic Mean and L2 Norm in comparison to other methods. In terms of classification accuracy, the performance of the aggregation methods was generally data-dependent with demonstrated significant differences in classification accuracy between score-based aggregations and rank-based aggregations. In terms of stability, the experimental results showed that both the Arithmetic Mean and L2 Norm result in better stability behaviour than any other aggregation rule. In fact, we discovered that these two score-based aggregations, for the most part, produced better results than any other more complex alternative techniques. In contribution to this field, this work considers the influence of the aggregation method in evaluating the stability of feature selection ensembles and underlines both the accuracy and stability differences between score-based and rank-based aggregations. However, this study is limited to the aforementioned aggregation strategies and to the tested binary and multiclass classification datasets. Thus, a generalization of the results requires more validation. Future investigations can introduce more score-based and rank-based aggregations to the comparison and extend the underlying feature selection methods to others including embedded and wrapper techniques. Such findings can have significant practical implications in

terms of the identification of the aggregation techniques that can be highly appropriate for certain application domains.

Abbreviations
WAM: Within Aggregation Method; BAM: Between Aggregation Method; IG: Information Gain; SU: Symmetric Uncertainty; CH: Chi-Square; MRMR: Minimum Redundancy Maximum Relevance; RRA: Robust Rank Aggregation; ANOVA: Analysis of Variance.

Acknowledgements
The authors are grateful for the comments and suggestions by the referees and the Editor. Their comments and suggestions have greatly improved the paper. The authors would like to express their great appreciation to the American University of Sharjah and the Second Forum for Women in Research Award for their support.

Author contributions
All authors have contributed to all sections including the methodology, the data analysis, and the conclusion. All authors read and approved the final manuscript.

Funding
This work was supported by the American University of Sharjah. The partial support from Emirates NBD and DEWA Research & Development Center via the Second Forum for Women in Research Award is also acknowledged. The second author would like to acknowledge the Faculty Research Grant # FRG21-S-S03.

Availability of data and materials
The datasets analysed during the current study are available from automl.chalearn.org, www.openml.org, mulan.sourceforge.net and archive.ics.uci.edu. Please see Table 1 for more details.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 October 2021 Accepted: 6 April 2022
Published online: 28 April 2022

References

1. Kumar V, Minz S. Feature selection: a literature review. SmartCR. 2014;4(3):211–29.
2. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3(Mar):1157–82.
3. Sulieman H, Alzaatreh A. A supervised feature selection approach based on global sensitivity. Arch Data Sci Ser A (Online First). 2018;5(1):03.
4. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
5. Venkatesh B, Anuradha J. A review of feature selection and its methods. Cybern Inf Technol. 2019;19(1):3–26.
6. Pes B. Evaluating feature selection robustness on high-dimensional data. In: International conference on hybrid artificial intelligence systems. Springer; 2018. p. 235–247.
7. Alelyani S. Stable bagging feature selection on medical data. J Big Data. 2021;8(1):1–18.
8. Brown G. Ensemble learning. Encycl Mach Learn. 2010;312:15–9.
9. Salman R, Alzaatreh A, Sulieman H, Faisal S. A bootstrap framework for aggregating within and between feature selection methods. Entropy. 2021;23(2):200.
10. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Joint European conference on machine learning and knowledge discovery in databases. Springer; 2008. p. 313–325.
11. Wang H, Khoshgoftaar TM, Napolitano A. A comparative study of ensemble feature selection techniques for software defect prediction. In: 2010 Ninth international conference on machine learning and applications. IEEE; 2010. p. 135–140.
12. Hoque N, Singh M, Bhattacharyya DK. Efs-mi: an ensemble feature selection method for classification. Complex Intell Syst. 2018;4(2):105–18.
13. Drotár P, Gazda M, Vokorokos L. Ensemble feature selection using election methods and ranker clustering. Inf Sci. 2019;480:365–80.
14. Chen C-W, Tsai Y-H, Chang F-R, Lin W-C. Ensemble feature selection in medical datasets: combining filter, wrapper, and embedded feature selection results. Expert Syst. 2020;37(5):12553.

15. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2009;26(3):392–8. https://doi.org/10.1093/bioinformatics/btp630. https://academic.oup.com/bioinformatics/article-pdf/26/3/392/16896736/btp630.pdf
16. Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput Appl. 2020;32(10):5951–73.
17. Liu H, Motoda H, Setiono R, Zhao Z. Feature selection: An ever evolving frontier in data mining. In: Liu, H., Motoda, H., Setiono, R., Zhao, Z. (eds.) Proceedings of the Fourth International Workshop on Feature Selection in Data Mining. Proceedings of Machine Learning Research, vol. 10, pp. 4–13. PMLR, Hyderabad, India (2010). https://proceedings.mlr.press/v10/liu10b.html.
18. Piramuthu S. Evaluating feature selection methods for learning in data mining applications. Eur J Oper Res. 2004;156(2):483–94.
19. Liu H, Motoda H. Computational methods of feature selection. Cham: CRC Press; 2007.
20. Guan D, Yuan W, Lee Y-K, Najeebullah K, Rasel MK. A review of ensemble learning based feature selection. IETE Tech Rev. 2014;31(3):190–8.
21. Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: a review and future trends. Inf Fusion. 2019;52:1–12.
22. Onan A, Korukoğlu S. A feature selection model based on genetic rank aggregation for text sentiment classification. J Inf Sci. 2017;43(1):25–38.
23. Najdi S, Gharbali AA, Fonseca JM. Feature ranking and rank aggregation for automatic sleep stage classification: a comparative study. Biomed Eng Online. 2017;16(1):1–19.
24. López-Cabrera JD, Lorenzo-Ginori JV. Feature selection for the classification of traced neurons. J Neurosci Methods. 2018;303:41–54.
25. Lin S. Rank aggregation methods. Wiley Interdiscip Rev Comput Stat. 2010;2(5):555–70.
26. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, et al. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006;24(5):537–44.
27. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012;28(4):573–80.
28. Joachims T. Optimizing search engines using clickthrough data. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. 2002. p. 133–142
29. Dittman DJ, Khoshgoftaar TM, Wald R, Napolitano A. Classification performance of rank aggregation techniques for ensemble gene selection. In: The twenty-sixth international FLAIRS conference 2013.
30. Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, Alonso-Betanzos A. Ensemble feature selection: homogeneous and heterogeneous approaches. Knowl Based Syst. 2017;118:124–39.
31. Seijo-Pardo B, Bolón-Canedo V, Alonso-Betanzos A. Using a feature selection ensemble on dna microarray datasets. In: ESANN 2016.
32. Seijo-Pardo B, Bolón-Canedo V, Alonso-Betanzos A. Testing different ensemble configurations for feature selection. Neural Process Lett. 2017;46(3):857–80.
33. Wald R, Khoshgoftaar TM, Dittman D, Awada W, Napolitano A. An extensive comparison of feature ranking aggregation techniques in bioinformatics. In: 2012 IEEE 13th international conference on information reuse and integration (IRI). IEEE; 2012. p. 377–384.
34. Wald R, Khoshgoftaar TM, Dittman D. Mean aggregation versus robust rank aggregation for ensemble gene selection. In: 2012 11th international conference on machine learning and applications, vol. 1. IEEE; 2012. p. 63–69.
35. Dessì N, Pes B, Angioni M. On stability of ensemble gene selection. In: International conference on intelligent data engineering and automated learning. Springer; 2015. p. 416–423.
36. Willett P. Combination of similarity rankings using data fusion. J Chem Inf Model. 2013;53(1):1–10.
37. Dittman DJ, Khoshgoftaar TM, Wald R, Napolitano A. Comparison of rank-based vs. score-based aggregation for ensemble gene selection. In: 2013 IEEE 14th international conference on information reuse and integration (IRI). IEEE; 2013. p. 225–231.
38. Dernoncourt D, Hanczar B, Zucker J-D. Stability of ensemble feature selection on high-dimension and low-sample size data. In: Proceedings of the 3rd international conference on pattern recognition applications and methods. 2014. p. 325–330.
39. Li Y, Hsu DF, Chung SM. Combining multiple feature selection methods for text categorization by using rank-score characteristics. In: 2009 21st IEEE international conference on tools with artificial intelligence. IEEE; 2009. p. 508–517.
40. Alelyani S, Zhao Z, Liu H. A dilemma in assessing stability of feature selection algorithms. In: 2011 IEEE international conference on high performance computing and communications. IEEE; 2011. p. 701–707.
41. Dittman D, Khoshgoftaar T, Wald R, Napolitano A. Similarity analysis of feature ranking techniques on imbalanced dna microarray datasets. In: 2012 IEEE international conference on bioinformatics and biomedicine. IEEE; 2012. p. 1–5.
42. Wald R, Khoshgoftaar TM, Napolitano A. Stability of filter-and wrapper-based feature subset selection. In: 2013 IEEE 25th international conference on tools with artificial intelligence. IEEE; 2013. p. 374–380.
43. Lustgarten JL, Gopalakrishnan V, Visweswaran S. Measuring stability of feature selection in biomedical datasets. In: AMIA annual symposium proceedings, vol. 2009. American Medical Informatics Association; 2009. p. 406.
44. Nogueira S, Brown G. Measuring the stability of feature selection with applications to ensemble methods. In: International workshop on multiple classifier systems. Springer; 2015. p. 135–146.
45. Kuncheva LI. A stability index for feature selection. In: Artificial intelligence and applications. 2007. p. 421–427.
46. Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. J Mach Learn Res. 2017;18(1):6345–98.
47. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst. 2007;12(1):95–116.
48. Bommert, A., Rahnenführer, J.: Adjusted measures for feature selection stability for data sets with similar features. In: International conference on machine learning, optimization, and data science. Springer; 2010. p. 203–214

49. Yu E, Cho S. Ensemble based on ga wrapper feature selection. Comput Ind Eng. 2006;51(1):111–6.
50. Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: a review. J King Saud Univ Comput Inf Sci 2019;34(4):1060–1073. https://doi.org/10.1016/j.jksuci.2019.06.012
51. Kent JT. Information gain and a general measure of correlation. Biometrika. 1983;70(1):163–73.
52. Muirhead R. Proofs that the arithmetic mean is greater than the geometric mean. Math Gaz. 1903;2(39):283–7.

Publisher's Note