


## Article

# Feature Ranking and Screening for Class-Imbalanced Metabolomics Data Based on Rank Aggregation Coupled with Re-Balance

Guang-Hui Fu <sup>1,\*</sup> , Jia-Bao Wang <sup>1</sup>, Min-Jie Zong <sup>1</sup> and Lun-Zhao Yi <sup>2,\*</sup>
<sup>1</sup> School of Science, Kunming University of Science and Technology, Kunming 650500, China; wangjiabao@126.com (J.-B.W.); zongminjie@126.com (M.-J.Z.)

<sup>2</sup> Faculty of Agriculture and Food, Kunming University of Science and Technology, Kunming 650500, China

\* Correspondence: guanghui fu@kust.edu.cn (G.-H.F.); yilunzhao@kust.edu.cn (L.-Z.Y.)



**Citation:** Fu, G.-H.; Wang, J.B.; Zong, M.-J.; Yi, L.-Z. Feature Ranking and Screening for Class-Imbalanced Metabolomics Data Based on Rank Aggregation Coupled with Re-Balance. *Metabolites* **2021**, *11*, 389. <https://doi.org/10.3390/metabo11060389>

Academic Editor: Hunter N. B. Moseley

Received: 9 April 2021

Accepted: 8 June 2021

Published: 14 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Feature screening is an important and challenging topic in current class-imbalance learning. Most of the existing feature screening algorithms in class-imbalance learning are based on filtering techniques. However, the variable rankings obtained by various filtering techniques are generally different, and this inconsistency among different variable ranking methods is usually ignored in practice. To address this problem, we propose a simple strategy called rank aggregation with re-balance (RAR) for finding key variables from class-imbalanced data. RAR fuses each rank to generate a synthetic rank that takes every ranking into account. The class-imbalanced data are modified via different re-sampling procedures, and RAR is performed in this balanced situation. Five class-imbalanced real datasets and their re-balanced ones are employed to test the RAR's performance, and RAR is compared with several popular feature screening methods. The result shows that RAR is highly competitive and almost better than single filtering screening in terms of several assessing metrics. Performing re-balanced pretreatment is hugely effective in rank aggregation when the data are class-imbalanced.

**Keywords:** class-imbalance; feature screening; rank aggregation; re-balance; filtering algorithm

## 1. Introduction

Datasets with imbalanced distribution are quite common in classification. In the settings of binary category, a dataset is called “imbalanced” if the number of one class is far larger than the others in the training data. Generally, the majority class is called negative while the minority class is called positive. Thus, the number of positive instances is often much lower than that of negative ones.

A hindrance in class-imbalance learning is that standard classifiers are often biased towards the majority classes. Therefore, there is a higher misclassification rate in the minority instances [1,2]. Re-sampling is the standard strategy to deal with class-imbalance learning tasks. Many studies [2–4] have shown that re-sampling the dataset is an effective way to enhance the overall performance of the classification for several types of classifiers. Re-sampling methods concentrate on modifying the training set to make it suitable for a standard classifier. There are generally three types of re-sampling strategies to balance the class distribution: over-sampling, under-sampling, and hybrid sampling.

- Over-sampling adds a set sampled from the minority class. Randomly duplicating the minority instances, SMOTE [5] and smoothed bootstrap [6] are three widely used over-sampling methods.
- Under-sampling removes some of the data points from the majority class to alleviate the harms of imbalanced distribution. Random under-sampling (RUS) is a simple but effective way to randomly remove part of the majority class.
- Hybrid-sampling is a combination of over-sampling and under-sampling.

Let  $D$  be a dataset with  $p$  features  $x_1, x_2, \dots, x_p$ , the target of feature screening is to extract a part of features  $x'_1, x'_2, \dots, x'_m$  such that  $m \ll p$  and these selected features satisfy the specified conditions of the task at hand [7]. For instance, the target is to select the subset of candidate features to maximize classifier accuracy in a classification setting. In the past two decades, many papers in studies have adopted the feature screening methods [8–10]. Feature screening has many advantages such as reducing susceptibility to over-fitting, training models faster and offsetting the pernicious effects of the curse of dimensionality [8]. The disadvantage of feature screening is that some crucial features may be omitted, thus harming classification performance.

Filtering [11], wrapping [12], and embedding [13] are three kinds of approaches for feature screening. Filter algorithms screen top-ranked variables via a certain metric. Wrapper methods perform a search in all the combinations to find the best subsets of all features. Generally, a complete search is often time-consuming and greedy, so the heuristic technique is frequently utilized to explore the solutions. Embedded algorithms screen important variables while building the classifier. Of all the three types of feature screening, filter methods are the simplest and the most frequently used to solve real-world imbalanced problems [14] in class-imbalance learning community. Many metrics have been utilized to perform filtering feature screening algorithms, such as  $t$  test, Fisher score [15], Hellinger distance [16], Relief [17], ReliefF [18], information gain [19], Gini index [20], AUCROC [21], AUCPRC [22], geometric mean [23], F-measure [24], and R-value [25].

Ensemble feature selection has been widely applied to the field of classification [26], such as Nazrul et al. [27] provided an ensemble feature selection method using feature-class and feature mutual information to select an optimal subset of features by combining multiple subsets of features. Yang et al. [28] proposed an ensemble-based wrapper approach for feature selection from data with highly imbalanced class distribution. Nowadays, feature selection methods are popular in metabolomics data analysis. In order to resolve the problem of filtering the discriminative metabolites from high-dimension metabolomics data, Lin et al. [29] proposed a mutual information (MI)-SVM-RFE method that filters out noise and non-informative variables by means of artificial variables and MI, then conducts SVM-RFE to select the most discriminative features. Fu et al. [30] proposed two feature selection algorithms that, by minimizing the overlap degree between the majority and the minority, are effective in recognizing key features and control false discoveries for class-imbalanced metabolomics data. The above feature screening methods are usually established for balanced datasets, but they are also directly utilized in class-imbalance situations.

Different filtered approaches give different feature rankings because of their different theories, even when just counting top-ranked features. Motivated by this problem, we propose a simple strategy called rank aggregation with re-balance (RAR) to combine all methods' ranking results in this study. It is an essential tool to fuse each rank to generate a synthetic rank that takes every ranking into account for class-imbalanced data. Different from the general feature selection methods, the proposed method combines different feature selection methods rather than simply accepting the result of one method, which can enhance the stability of the algorithm. At the same time, the great performances of the experiments in balanced and imbalanced metabolomics datasets verify the strong generalization abilities of RAR.

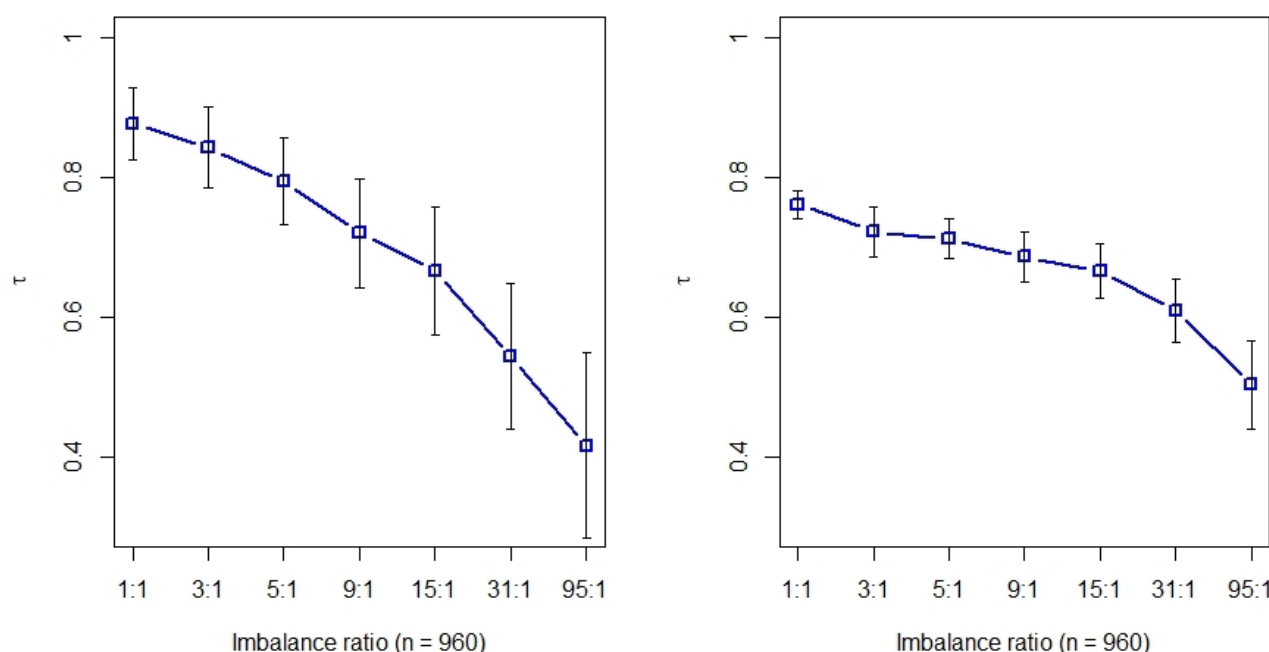
## 2. Results

### 2.1. Kendall's $\tau$ Rank Correlation of Eight Filtering Methods on Class-Imbalanced Data

Each filtered method above can be employed to perform feature screening. However, we noted that different filtering feature screening techniques may give different rankings, especially when the data are extremely class-imbalanced. In this section, we compare methods using Kendall's  $\tau$  rank correlation [31].

The Kendall's  $\tau$  rank correlation of eight filtering methods ( $t$  test, Fisher score, Hellinger distance, Relief, ReliefF, Information gain, Gini index, and R-value) are computed with simu-

lated data that are generated by multivariate normal distributions, namely,  $\mathbf{X}|(y = 0) \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$  and  $\mathbf{X}|(y = 1) \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ , where the label  $y = 0$  denotes the majority class and  $y = 1$  minority class, respectively. The predictors in two classes have the same covariance matrix  $\boldsymbol{\Sigma}$ , which is set to be a unit matrix for the purpose of simplicity. Two cases are considered in this study. In case one, the number of  $p = 8$ , and eight variables are all set to be key features. The difference of mean values  $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = [2.4, 2.2, 2, 1.8, 1.6, 1.4, 1.2, 1]$ . In case two,  $p = 16$ , and the first eight variables are set to be the same with case one, but another eight irrelevant predictors are added. The number of total instances is set to 960. The negative to the positive ratios here are set to be 1:1, 3:1, 9:1, 31:1, and 95:1, respectively. There are 28 Kendall's  $\tau$  rank correlation coefficients among 8 filtering methods, and the mean of these coefficients (with 100 repeats) is shown in Figure 1. As stated above,  $\tau = 1$  if all pairs are concordant. Whereas the maximum of  $\tau$  is 0.88 in case one (left, Figure 1) where there are no irrelevant predictors, and 0.76 in case two (right, Figure 1) where one-half of features are irrelevant variables. The two maximal  $\tau$  values are reached when two classes are exactly balanced, and  $\tau$  reduces as the imbalance ratio increases in two cases. It indicates that these filtering methods probably generate different feature rankings, and such differences tend to be intensified when the class imbalanced ratio increases. Consequently, it is hard to say that a filtering approach is better or worse than another one, and it is a big risk to just depend on a single filter algorithm to make decisions. We have known that such a difference will occur due to the different principles of the filtering methods, but we also presume that class imbalance intensifies this difference. A natural way to combat this challenge may combine each filtering approach's information and relieve the effect of class imbalance. This is the motivation for why we propose the strategy of rank aggregation with re-balance.

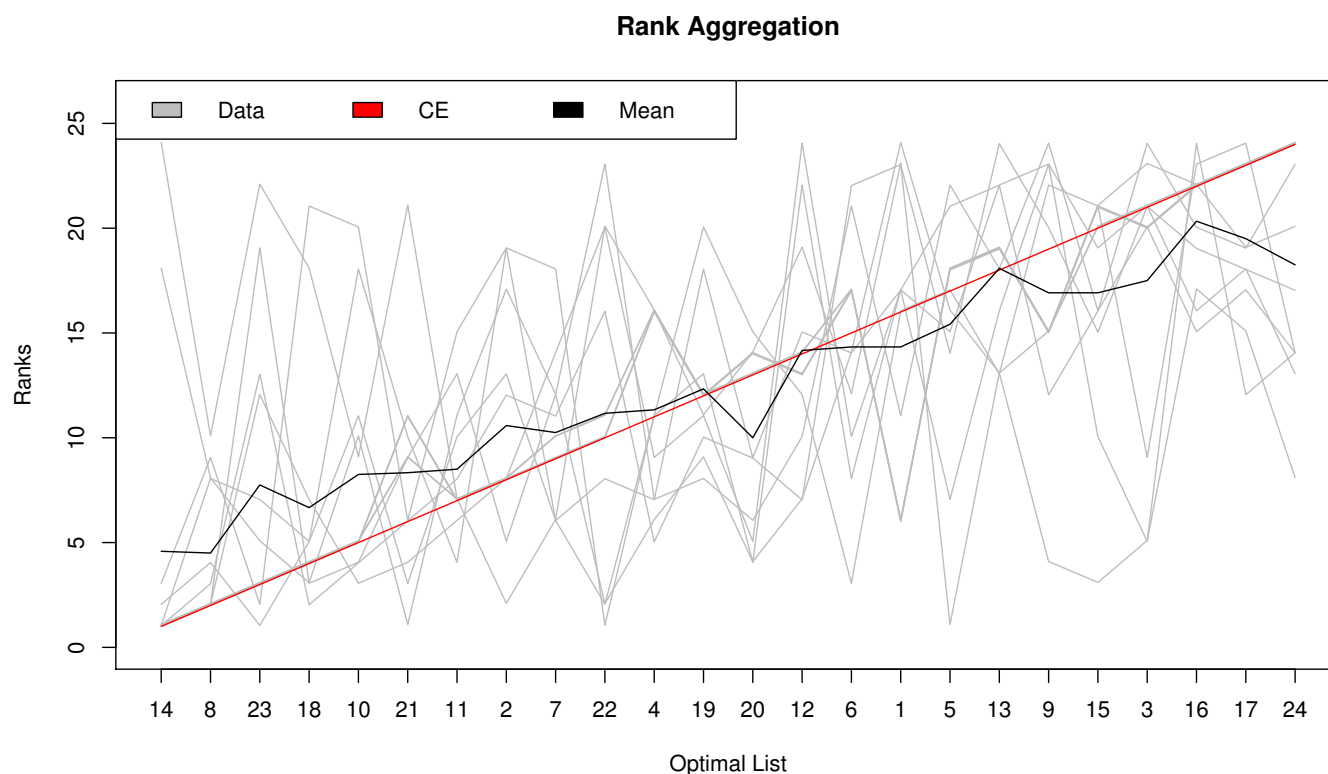


**Figure 1.** Kendall's  $\tau$  rank correlation coefficient under different imbalance ratios (with 100 repeats). **Left:** eight key variables; **right:** eight key plus eight irrelevant variables.

## 2.2. Rank Aggregation(RA) on Original Balanced Data

In our computation, eight filtering methods— $t$  test, Fisher score, Hellinger distance, Relief, ReliefF, Information gain, Gini index, and R-value—are aggregated to generate an incorporative rank. Rank aggregation is firstly tested with the original balanced dataset "NPC". Artificial rebalancing is unnecessary, and just case 1 (no resampling) is performed. Rank aggregation is compared with eight filtering methods:  $t$  test, Fisher score, Hellinger distance, Relief, ReliefF, Information gain, Gini index, and R-value.  $G_{mean}$ ,  $F_1$ ,  $AUCROC$

and *AUCPRC* are utilized as evaluation measurements. The rank lists ordered by their importance are shown on the *x*-axis in Figure 2. The top seven features are selected according to all four assessment metrics.



**Figure 2.** Rank lists of rank aggregation with the dataset “NPC” in case 1.

### 2.3. Rank Aggregation with Re-Balance (RAR) on Imbalanced Data

Figures 3–6 show the aggregated rank lists on seven cases with the datasets “TBI”, “CHD2-1”, “CHD2-2”, and “ATR”, respectively. Rank aggregation combines each ranking into a list reflective of the overall preference, and each subgraph of four figures shows the aggregation results based on the CE algorithm. The *x*-axis is the optimal list obtained by the rank aggregation algorithm. The *y*-axis also ranks, and the gray line is the rank of the original data; the black line is their average rank; and the red line is the aggregate result of the CE algorithm. The order of the *x*-axis rank is based on the aggregate ranks obtained by the red line. The performances measured by *Gmean*, *F<sub>1</sub>*, *AUCROC*, and *AUCPRC* are given in Tables 1–4, respectively.



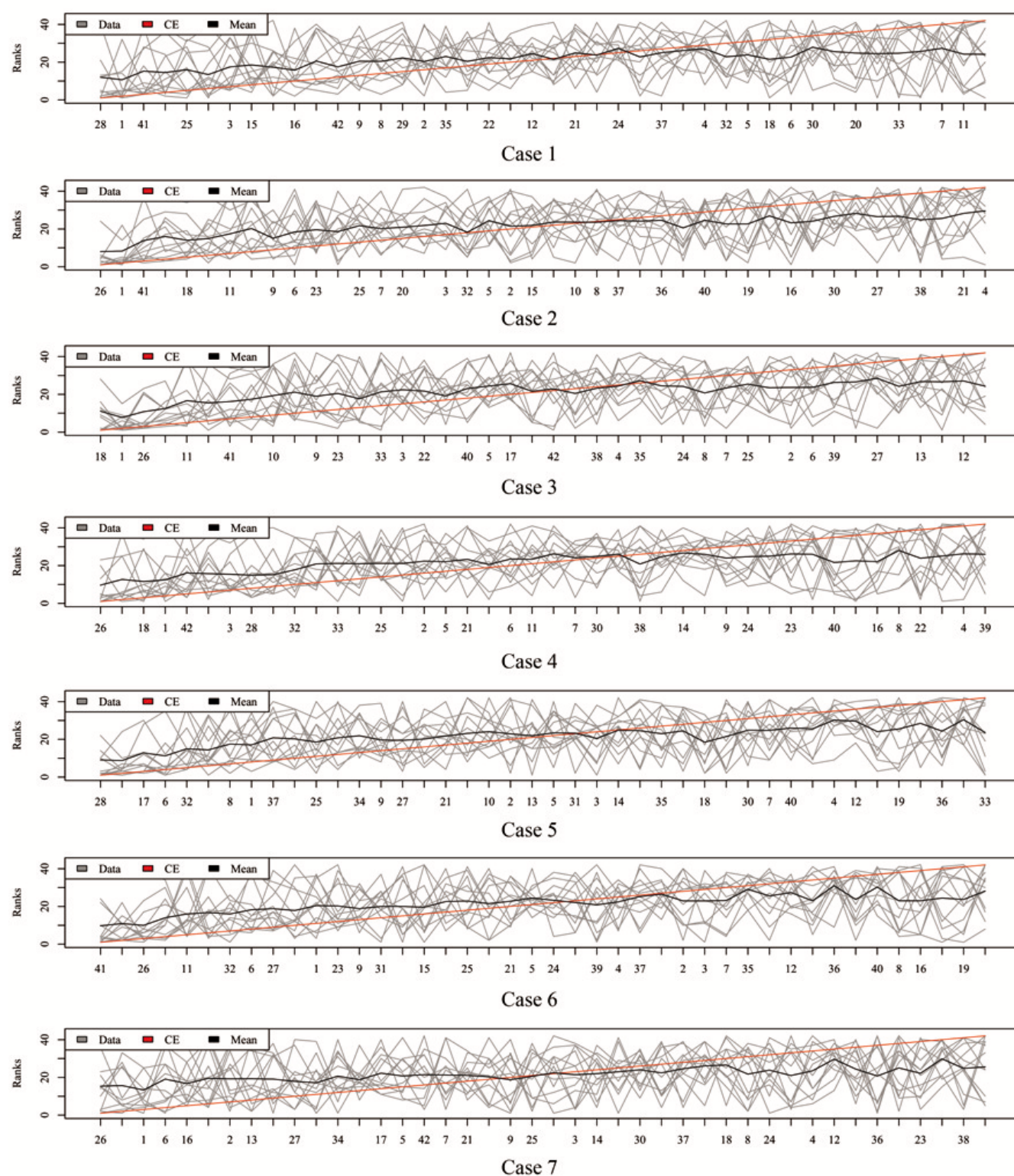


Figure 3. Rank lists of rank aggregation with the dataset “TBI” on seven cases.

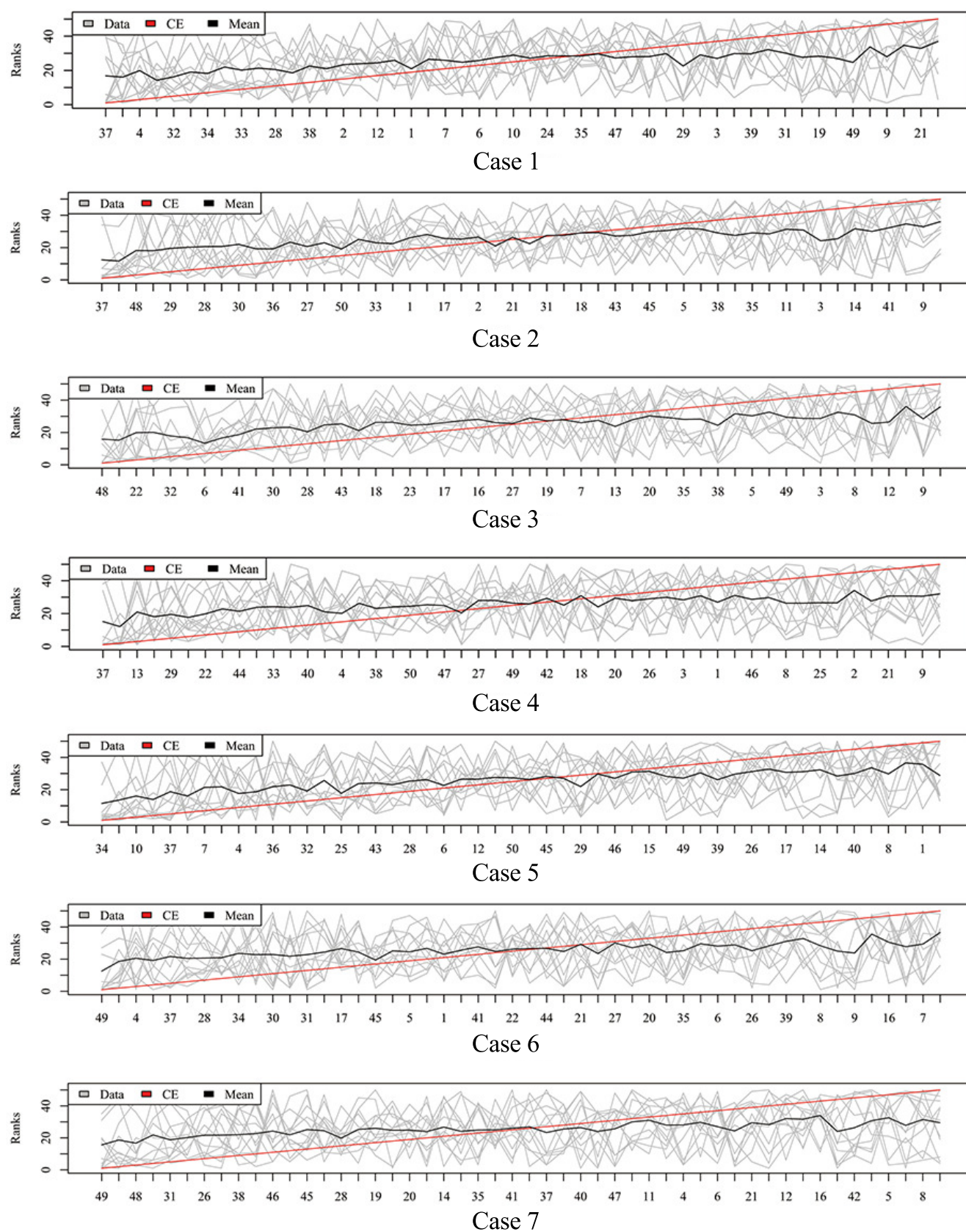
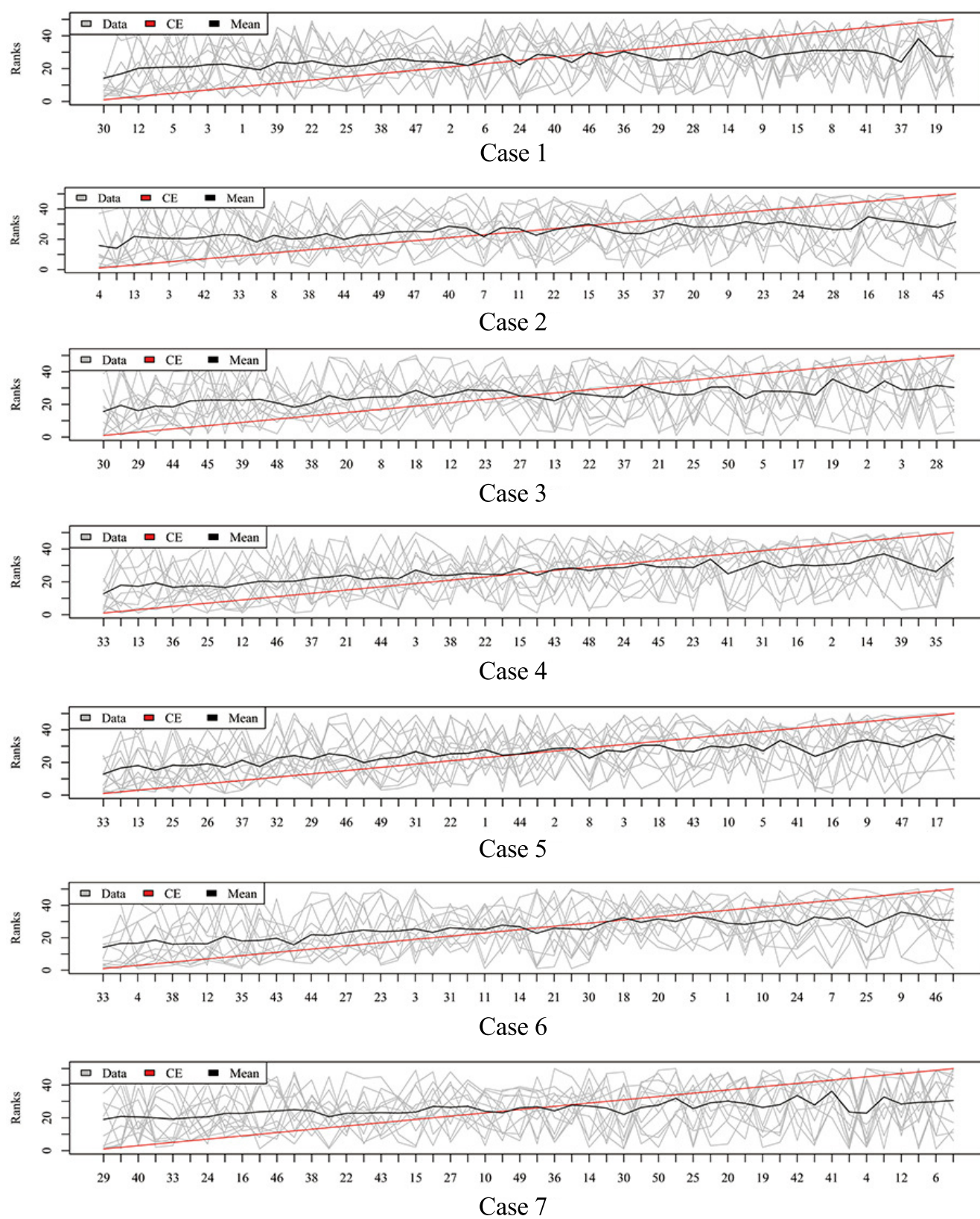


Figure 4. Rank lists of rank aggregation with the dataset "CHD2-1" on seven cases.





**Figure 5.** Rank lists of rank aggregation with the dataset "CHD2-2" on seven cases.

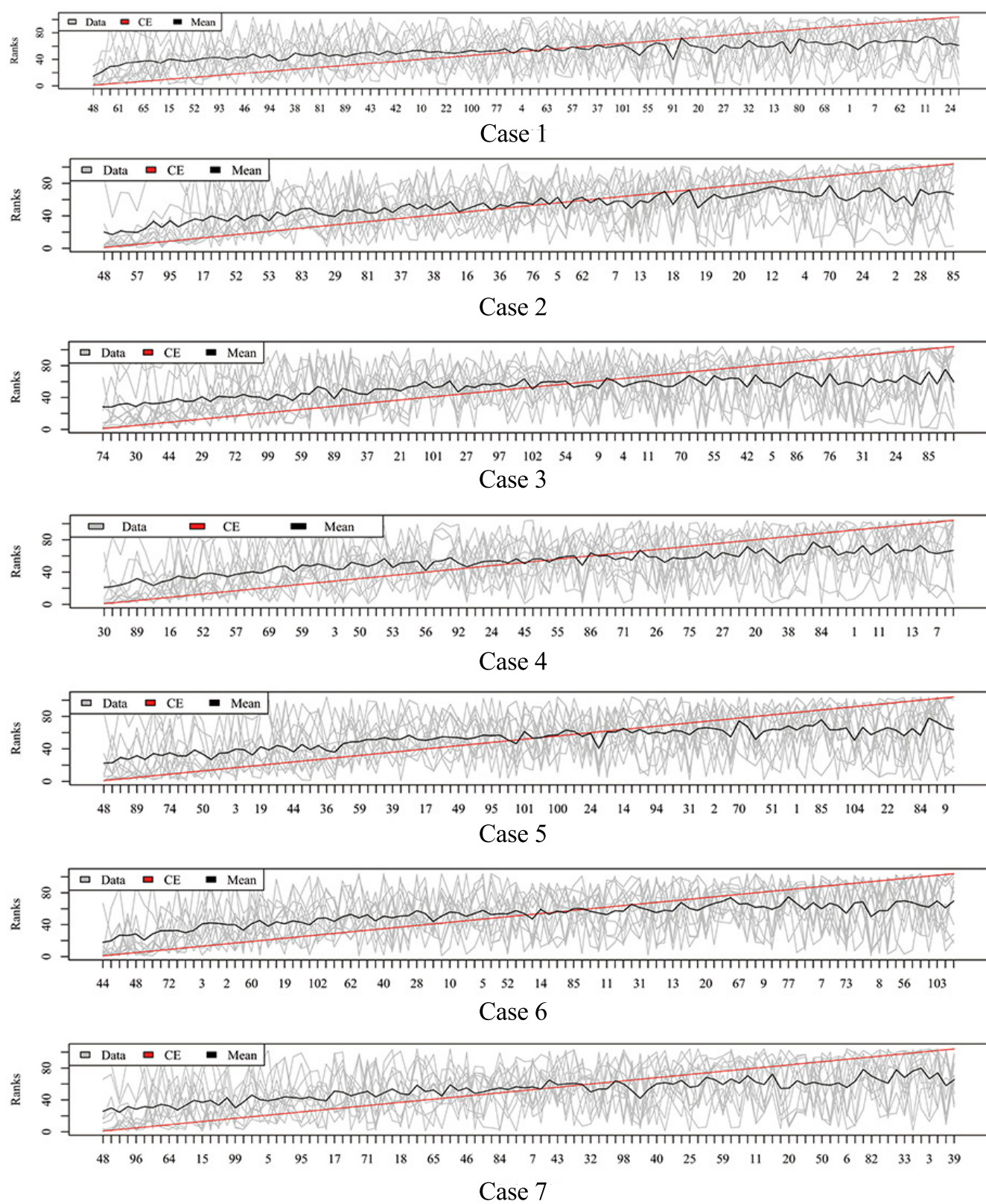


Figure 6. Rank lists of rank aggregation with the dataset "ATR" on seven cases.



**Table 1.** *Gmean* from rank aggregation and eight filtering techniques (The best result is in bold).

Dataset	Resampling	No.	RA	RAR	<i>t</i> Test	Fisher	Hellinger	Relief	ReliefF	IG	Gini	R-Value
NPC	Case 1	7	<b>1.00</b>	—	0.87	0.97	0.95	0.95	0.95	0.95	<b>1.00</b>	0.92
	TBI	6	<b>0.96</b>	—	0.41	0.68	0.68	0.70	0.88	0.72	0.58	0.63
CHD2-1	Case 2	11	—	<b>1.00</b>	0.84	0.88	0.95	0.90	0.94	0.86	0.90	0.90
	Case 3	1	—	<b>1.00</b>	0.95	0.84	0.95	0.90	0.80	0.95	0.90	<b>1.00</b>
	Case 4	10	—	<b>1.00</b>	0.96	<b>1.00</b>	0.89	0.93	0.97	0.93	0.85	<b>1.00</b>
	Case 5	7	—	<b>1.00</b>	0.93	0.90	0.97	0.97	0.93	0.97	0.86	0.93
	Case 6	12	—	<b>1.00</b>	0.75	0.83	0.82	0.83	0.91	0.85	0.71	<b>1.00</b>
	Case 7	29	—	<b>1.00</b>	0.71	0.70	0.65	0.82	0.85	0.78	0.71	0.71
	Case 1	10	<b>0.87</b>	—	0.67	0.71	<b>0.87</b>	0.00	0.77	0.77	0.47	<b>0.87</b>
	Case 2	11	—	0.94	0.85	0.85	<b>1.00</b>	0.87	0.93	0.94	0.85	0.91
	Case 3	6	—	<b>1.00</b>	0.86	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.93	<b>1.00</b>	0.93	<b>1.00</b>
	Case 4	31	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.95	<b>1.00</b>
	Case 5	37	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 6	11	—	0.87	0.61	0.75	0.87	0.87	<b>0.89</b>	0.87	0.50	0.87
	Case 7	10	—	<b>0.87</b>	0.61	0.71	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>	0.71	0.71
CHD2-2	Case 1	3	0.58	—	0.48	0.58	0.48	0.58	0.55	0.68	0.00	<b>0.73</b>
	Case 2	34	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 3	14	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 4	24	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 5	28	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 6	7	—	<b>1.00</b>	<b>1.00</b>	0.82	0.82	<b>1.00</b>	0.82	0.82	0.47	0.82
	Case 7	4	—	0.86	0.82	<b>1.00</b>	0.82	<b>1.00</b>	0.61	<b>1.00</b>	0.00	0.75
ATR	Case 1	25	<b>1.00</b>	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.50	<b>1.00</b>
	Case 2	9	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 3	8	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 4	2	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.89	<b>1.00</b>
	Case 5	2	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.80	<b>1.00</b>
	Case 6	9	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.71	<b>1.00</b>
	Case 7	10	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>

**Table 2.** *F<sub>1</sub>* from rank aggregation and eight filtering techniques (The best result is in bold).

Dataset	Resampling	NO.	RA	RAR	<i>t</i> Test	Fisher	Hellinger	Relief	ReliefF	IG	Gini	R-Value
NPC	Case 1	8	<b>1.00</b>	—	0.88	0.95	0.95	0.97	<b>1.00</b>	0.95	0.95	0.92
	TBI	11	<b>0.93</b>	—	0.82	0.86	0.87	0.90	0.90	0.88	0.83	0.88
CHD2-1	Case 2	13	—	<b>1.00</b>	0.84	0.90	0.94	0.95	0.94	0.84	0.87	0.86
	Case 3	1	—	<b>1.00</b>	0.87	0.87	<b>1.00</b>	<b>1.00</b>	0.96	0.95	0.83	<b>1.00</b>
	Case 4	7	—	<b>1.00</b>	0.91	0.97	0.93	0.97	0.90	0.93	0.89	<b>1.00</b>
	Case 5	10	—	<b>1.00</b>	0.87	0.93	0.93	0.93	0.93	0.97	0.88	0.93
	Case 6	19	—	<b>1.00</b>	0.73	0.80	0.86	0.86	0.77	0.80	0.71	0.91
	Case 7	11	—	<b>1.00</b>	0.83	0.92	0.80	0.86	0.86	0.75	0.50	0.86
	Case 1	10	<b>0.88</b>	—	0.87	<b>0.88</b>	0.87	0.87	0.87	0.87	0.83	0.87
	Case 2	11	—	<b>0.94</b>	<b>0.94</b>	0.71	0.93	0.89	0.93	0.88	0.67	0.89
	Case 3	6	—	<b>1.00</b>	0.89	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.93	0.94	0.94	0.92
	Case 4	31	—	<b>1.00</b>	0.95	0.95	0.96	0.90	0.96	0.90	0.86	0.95
	Case 5	37	—	<b>1.00</b>	<b>1.00</b>	0.91	0.95	0.95	<b>1.00</b>	0.95	0.86	0.95
	Case 6	11	—	0.89	0.50	0.75	0.73	0.89	<b>1.00</b>	0.83	0.55	0.67
	Case 7	10	—	<b>0.86</b>	0.73	0.57	0.80	0.75	<b>0.86</b>	0.67	0.55	0.57

Table 2. Cont.

Dataset	Resampling	NO.	RA	RAR	<i>t</i> Test	Fisher	Hellinger	Relief	ReliefF	IG	Gini	R-Value
CHD2-2	Case 1	3	0.91	—	0.87	0.86	<b>0.95</b>	0.87	0.91	0.90	0.87	0.91
	Case 2	34	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.88	0.92	0.93	<b>1.00</b>	0.88	0.93
	Case 3	14	—	<b>1.00</b>	<b>1.00</b>	0.92	0.86	0.92	<b>1.00</b>	0.93	0.86	0.93
	Case 4	24	—	<b>1.00</b>	0.90	0.91	0.95	0.95	0.95	<b>1.00</b>	0.95	0.95
	Case 5	28	—	<b>1.00</b>	0.95	0.95	0.95	0.95	<b>1.00</b>	0.96	0.95	<b>1.00</b>
	Case 6	7	—	<b>0.86</b>	0.57	0.80	0.80	0.75	<b>0.86</b>	<b>0.86</b>	0.50	<b>0.86</b>
	Case 7	4	—	<b>0.86</b>	0.80	0.75	0.57	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>	0.50	0.67
ATR	Case 1	25	<b>1.00</b>	—	<b>1.00</b>	0.89	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.89	<b>1.00</b>
	Case 2	9	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 3	8	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 4	2	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>0.89</b>	<b>1.00</b>
	Case 5	2	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
	Case 6	9	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>0.67</b>	<b>1.00</b>
	Case 7	10	—	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>0.67</b>	<b>1.00</b>

Table 3. AUCROC from rank aggregation and eight filtering techniques (The best result is in bold).

Dataset	Resampling	NO.	RA	RAR	<i>t</i> Test	Fisher	Hellinger	Relief	ReliefF	IG	Gini	R-Value
NPC	Case 1	16	<b>0.96</b>	—	0.90	0.91	0.94	0.93	0.96	0.93	0.93	0.95
	Case 2	3	<b>0.70</b>	—	0.48	0.61	0.52	0.61	0.65	0.58	0.49	0.67
TBI	Case 1	11	—	<b>0.89</b>	0.80	0.81	0.78	0.82	0.87	0.79	0.72	0.83
	Case 2	1	—	<b>0.95</b>	0.83	0.74	0.94	0.85	0.83	0.85	0.76	<b>0.95</b>
	Case 3	28	—	<b>0.91</b>	0.85	0.86	0.87	0.88	0.86	0.86	0.84	0.89
	Case 4	26	—	<b>0.93</b>	0.91	0.92	0.90	0.90	0.92	0.90	0.88	0.90
	Case 5	22	—	<b>0.77</b>	0.65	0.68	0.69	0.73	0.74	0.73	0.50	0.71
	Case 6	25	—	<b>0.71</b>	0.63	0.56	0.68	0.53	0.61	0.61	0.35	0.66
	Case 7	10	0.60	—	0.48	0.51	<b>0.61</b>	0.50	0.59	0.60	0.45	0.59
CHD2-1	Case 1	11	—	0.76	0.66	0.60	0.75	<b>0.77</b>	0.73	0.72	0.61	0.73
	Case 2	6	—	<b>0.92</b>	0.85	0.81	<b>0.92</b>	0.90	0.81	0.88	0.80	0.79
	Case 3	31	—	0.86	0.81	0.79	0.86	<b>0.87</b>	0.82	0.84	0.75	0.84
	Case 4	37	—	<b>0.87</b>	0.85	0.84	0.86	0.86	0.81	0.86	<b>0.90</b>	0.82
	Case 5	11	—	<b>0.64</b>	0.57	0.45	0.50	0.57	0.57	<b>0.64</b>	0.36	0.62
	Case 6	10	—	<b>0.60</b>	0.52	0.55	0.52	0.48	0.52	<b>0.60</b>	0.36	0.52
	Case 7	4	—	<b>0.66</b>	0.56	0.63	0.50	<b>0.66</b>	<b>0.66</b>	0.63	0.22	0.56
CHD2-2	Case 1	3	<b>0.60</b>	—	0.52	0.49	0.54	0.50	0.52	0.55	0.39	0.57
	Case 2	34	—	0.86	0.79	0.84	<b>0.88</b>	0.73	0.80	0.80	0.70	0.79
	Case 3	14	—	<b>0.90</b>	0.85	0.84	0.82	<b>0.90</b>	0.68	0.88	0.81	0.84
	Case 4	24	—	<b>0.93</b>	0.88	0.88	0.88	0.90	0.86	0.91	0.87	<b>0.93</b>
	Case 5	28	—	<b>0.91</b>	0.88	0.90	0.88	0.85	<b>0.91</b>	0.89	0.89	0.90
	Case 6	7	—	0.63	0.53	0.56	<b>0.66</b>	0.59	0.44	0.56	0.28	0.59
	Case 7	4	—	<b>0.66</b>	0.56	0.63	0.50	<b>0.66</b>	<b>0.66</b>	0.63	0.22	0.56
ATR	Case 1	25	0.88	<b>0.98</b>	—	0.51	0.85	0.85	0.76	0.85	0.50	0.76
	Case 2	9	—	<b>0.96</b>	0.79	0.86	<b>0.96</b>	<b>0.96</b>	0.93	<b>0.96</b>	0.75	0.93
	Case 3	8	—	<b>0.97</b>	<b>1.00</b>	0.90	0.97	0.90	0.97	0.97	0.80	0.93
	Case 4	2	—	<b>0.98</b>	0.93	0.83	0.88	<b>0.98</b>	0.95	<b>0.98</b>	0.81	0.95
	Case 5	2	—	<b>0.98</b>	0.81	0.86	0.76	0.93	0.93	0.95	0.71	0.88
	Case 6	9	—	<b>0.94</b>	0.81	0.75	0.88	<b>0.94</b>	0.88	0.81	0.19	0.81
	Case 7	10	—	<b>0.94</b>	0.81	0.63	0.69	0.63	<b>0.94</b>	0.88	0.19	0.75

**Table 4.** *AUCPRC* from rank aggregation and eight filtering techniques (The best result is in bold).

Dataset	Resampling	NO.	RA	RAR	<i>t</i> Test	Fisher	Hellinger	Relief	Relieff	IG	Gini	R-Value
NPC	Case 1	15	<b>0.96</b>	—	0.87	0.90	0.91	0.92	0.96	0.92	0.91	0.91
	TBI	8	<b>0.62</b>	—	0.27	0.58	0.40	0.47	0.56	0.47	0.26	0.46
TBI	Case 2	18	—	0.85	0.79	0.81	0.78	<b>0.86</b>	0.84	0.81	0.65	0.75
	Case 3	1	—	<b>0.93</b>	0.77	0.60	0.91	0.85	0.74	0.85	0.74	0.89
	Case 4	27	—	<b>0.89</b>	0.84	0.81	0.83	0.85	0.86	0.83	0.81	0.85
	Case 5	13	—	<b>0.91</b>	0.81	0.82	0.81	0.82	0.90	0.86	0.79	0.81
	Case 6	30	—	<b>0.78</b>	0.61	0.69	0.64	0.62	0.66	0.67	0.50	0.70
	Case 7	24	—	<b>0.71</b>	0.65	0.54	0.64	0.59	0.57	0.62	0.41	0.58
	Case 8	10	<b>0.61</b>	—	0.33	0.30	0.52	0.28	0.45	0.43	0.23	0.45
CHD2-1	Case 2	11	—	<b>0.77</b>	0.56	0.58	0.69	0.65	0.64	0.68	0.55	0.65
	Case 3	6	—	<b>0.88</b>	0.79	<b>0.88</b>	0.81	0.77	0.81	0.86	0.79	0.86
	Case 4	31	—	<b>0.86</b>	0.81	0.80	0.80	0.82	0.83	0.85	0.73	0.80
	Case 5	37	—	<b>0.87</b>	0.81	0.75	0.84	0.85	0.82	0.84	0.78	0.82
	Case 6	11	—	<b>0.66</b>	0.46	0.45	0.48	0.65	0.53	0.55	0.40	0.50
	Case 7	10	—	<b>0.63</b>	0.44	0.53	0.52	0.55	0.56	0.56	0.40	0.48
	Case 8	3	<b>0.45</b>	—	0.22	0.33	0.27	0.32	0.31	0.27	0.21	0.26
CHD2-2	Case 2	34	—	<b>0.87</b>	0.68	0.78	0.76	0.82	0.85	0.86	0.73	0.80
	Case 3	14	—	<b>0.87</b>	0.81	0.74	0.81	0.86	0.76	0.83	0.79	0.78
	Case 4	24	—	<b>0.91</b>	0.84	0.90	0.87	0.85	0.85	0.87	0.85	0.81
	Case 5	28	—	<b>0.90</b>	0.86	0.88	0.80	0.88	0.86	0.89	0.79	0.79
	Case 6	7	—	<b>0.71</b>	0.42	0.62	0.64	0.66	0.45	0.57	0.35	0.63
	Case 7	4	—	0.64	0.64	0.57	0.50	0.63	0.55	<b>0.66</b>	0.40	0.60
	Case 8	25	<b>0.93</b>	—	0.75	0.52	0.52	0.82	0.82	0.82	0.25	0.60
ATR	Case 2	9	—	<b>1.00</b>	0.81	0.88	0.92	0.94	0.94	0.94	0.68	0.88
	Case 3	8	—	<b>1.00</b>	<b>1.00</b>	0.78	<b>1.00</b>	0.88	0.93	<b>1.00</b>	0.82	0.88
	Case 4	2	—	<b>0.95</b>	0.84	0.79	0.91	0.91	0.88	0.91	0.70	0.91
	Case 5	2	—	<b>0.95</b>	0.79	0.78	0.78	0.88	0.90	<b>0.95</b>	0.61	0.88
	Case 6	9	—	<b>1.00</b>	0.89	0.69	0.85	0.89	0.76	0.85	0.36	0.80
	Case 7	10	—	<b>1.00</b>	0.61	0.54	0.64	0.54	0.80	0.89	0.37	0.80
	Case 8	10	—	<b>1.00</b>	0.61	0.54	0.64	0.54	0.80	0.89	0.37	0.80

### 3. Discussion

Tables 1 and 2 show that RA reached the maximal values of *Gmean* and  $F_1$ . It can be seen from Tables 3 and 4 that RA and Relieff obtained the maximal values of *AUCROC* and *AUCPRC*. Therefore, RA outperformed single filtering methods when assessed with *Gmean*,  $F_1$ , *AUCROC*, and *AUCPRC*. The NPC dataset had a completely balanced distribution, and RA worked well on it. Thus, rank aggregation is necessary to integrate different results, even if in a totally balanced situation, and a consensual feature ranking list is provided.

Aggregation ranking lists in Figures 3–6 tell us the order of importance of each feature. Though the rank lists derived from different subsampling methods were not the same, the top features were approximately consistent. After obtaining the rank list, another task is to figure out how many features should be considered as key variables. In this computation, we performed 5-fold cross-validation [32] to find the optimal number of key features. As recent studies have showed that *AUCPRC* is more informative in imbalanced learning [32,33], *AUCPRC* was employed as performance metric in this section, and random forest classifier was utilized to implement classification. Namely, the value of *AUCPRC* was calculated, as the top *k* ranked features were used each time, where *k* varies from 1 to *p* (see Figures 7–10). We chose the optimal *k*-value such that the random forest classifier had the maximal *AUCPRC* in identifying classification. It can be seen from Tables 1–4 that the optimal number of important features varied greatly under different re-balanced strategies. One possible reason is that the artificial data generated by different subsampling have difference to some extent. Another possible reason is the measurement changes slightly as the number of candidate features changes. This seems to be true from the Figures 7–10 where each curve tends to be flat as the changes in the number of features



used for classification. It also noted that the *AUCPRC* under no re-sampling (case 1) was generally lower than that under six re-sampling methods.

Tables 1–4 report the results of these real datasets with the assessing metrics *Gmean*,  $F_1$ , *AUCROC*, and *AUCPRC*, respectively. We can perform comparisons in several aspects. Original imbalanced datasets are employed in case 1 from Tables 1–4 (except NPC dataset). Of all the 16 “no re-balance” situations, the aggregation rank method reached the maximal measures in 12 situations compared with the other 8 filtering methods (t test, Fisher score, Hellinger distance, Relief, ReliefF, information gain, Gini, and R-value). It indicates that aggregation rank was better than a single filtering rank with the proportion of 75.00% when the data are class-imbalanced. If the original dataset NPC was counted in, this proportion was 80.00%. Therefore, rank aggregation is generally superior to single filtering methods, no matter how the data are balanced or imbalanced.

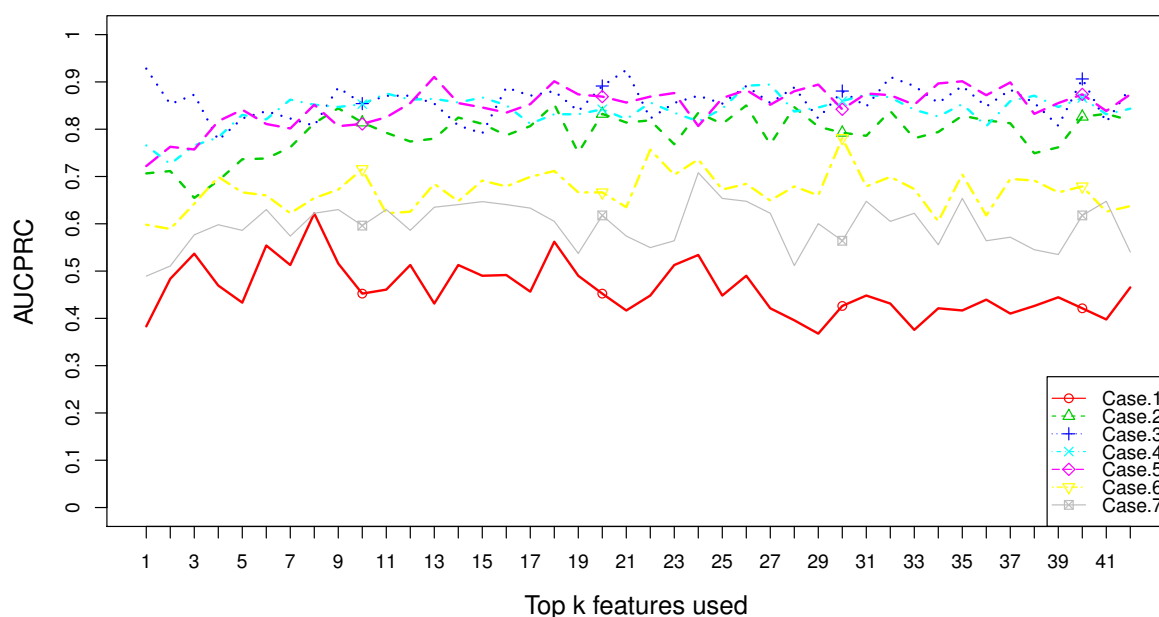


Figure 7. *AUCPRC* when top k features are used with the dataset “TBI” on seven cases.

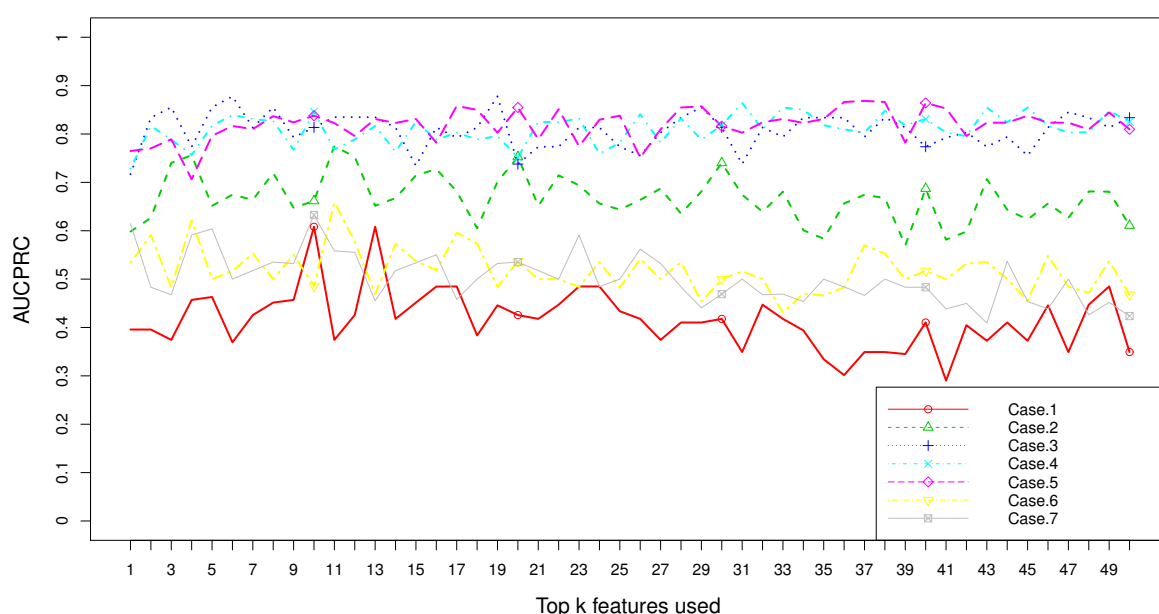
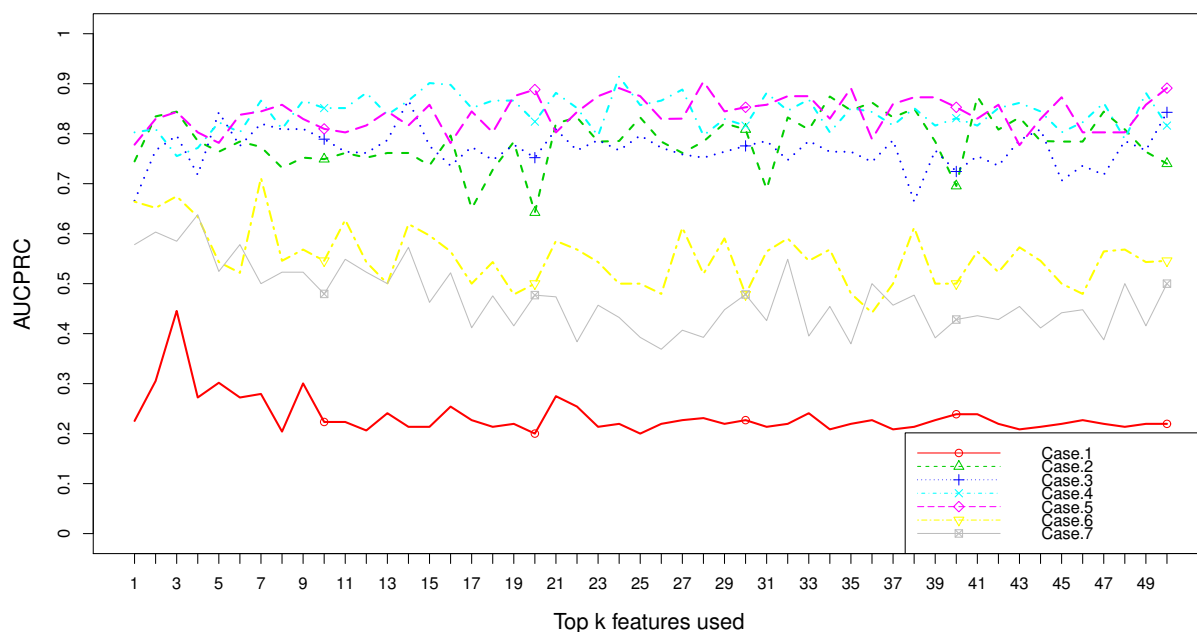
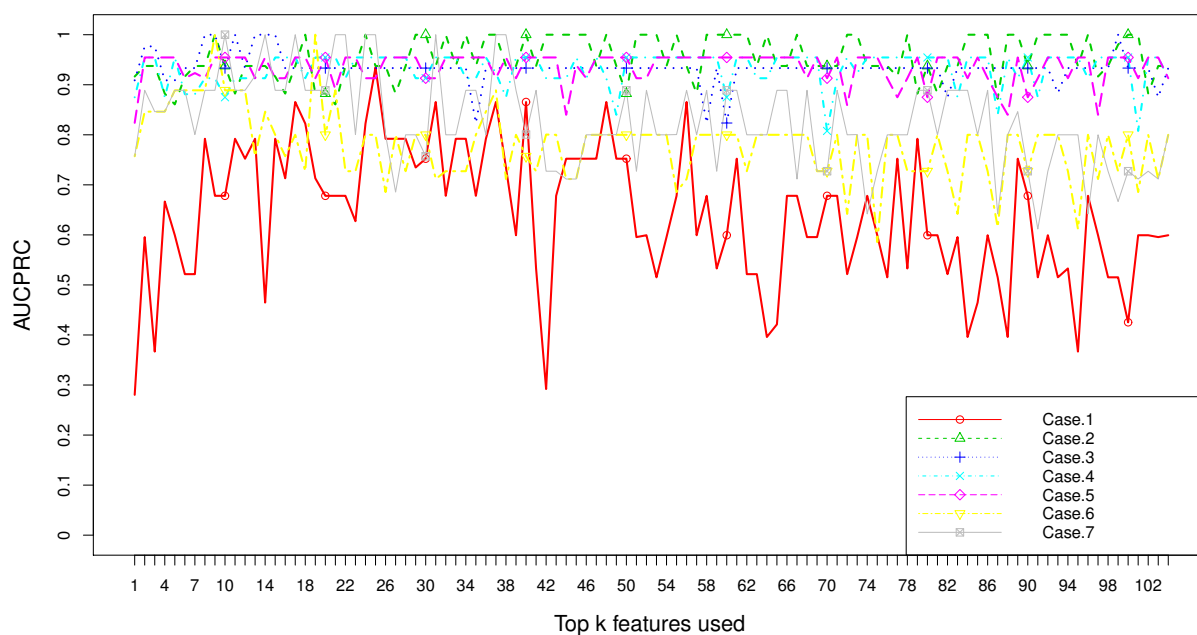


Figure 8. *AUCPRC* when top k features are used with the dataset “CHD2-1” on seven cases.



**Figure 9.** AUCPRC when top k features are used with the dataset “CHD2-2” on seven cases.



**Figure 10.** AUCPRC when top k features are used with the dataset “ATR” on seven cases.

Re-balanced datasets were artificially generated and utilized in cases 2–7 from Tables 1–4. Of all the 96 scenarios with re-balance, the aggregation rank method reached the maximal measures in 83 scenarios compared to the other eight filtering methods. It means that aggregation rank outperformed single filtering rank with the proportion of 86.46% when the class-imbalanced data were treated with re-balance strategies. Thus, performing aggregation rank is extremely effective in dealing with class-imbalanced data.

Rank aggregation was performed on both imbalanced datasets (RA) and re-balanced datasets (RAR). Of all the 96 scenarios with re-balance (cases 2–7 in Tables 1–4), there were 93 situations whose measurements were equal or greater than those from case 1 (no re-balance). It shows that aggregation rank with re-balance strategies performed better with the proportion of 96.88% than that with original class-imbalanced data. Therefore, perform-

ing re-balance can play a crucial role in improving the performance of rank aggregation when the data are class-imbalanced.

Figures 7–10 show the *AUCPRC* curves of seven cases on four imbalanced datasets. *AUCPRC* from re-balanced data (cases 2–7) was generally higher than that from imbalanced data (case 1). In other words, the performance can be promoted after re-sampling to balance the imbalanced data artificially. Case 5 and case 6 are two under-sampling methods, and the *AUCPRC* was generally lower than that from over-sampling or hybrid sampling (cases 2–4). The possible reason is that some of the useful information is missed in doing under-sampling when the size of the minority instances is too small (see Table 5). Therefore, one should be cautious about using under-sampling in practice.

**Table 5.** The summary of five datasets.

Datasets	Attributes	Instances	Majority	Minority	Ratio
NPC	24	200	100	100	1.00
TBI	42	104	73	31	2.35
CHD2-1	50	72	51	21	2.43
CHD2-2	50	67	51	16	3.19
ATR	104	29	21	8	2.63

In sum, different filter methods generate different rankings. Rank aggregation is necessary to integrate different results and provide a consensual feature ranking list. Class-imbalance usually leads to degraded performance from a filtering method on feature importance ranking. This harmfulness can be alleviated via different re-balance strategies in sample space.

## 4. Materials and Methods

### 4.1. Notations

The notations used in this study are listed below:

- $D$  a dataset with two classes  $C_1$  and  $C_2$
- $C_1$  the minority (positive) class
- $C_2$  the majority (negative) class
- $n$  the size of the total instances in  $D$
- $p$  the number of the features in  $D$
- $n_k$  the size of  $C_k$ ,  $k = 1, 2$
- $|D|$  the number of samples in  $D$
- $x_j$  the  $j$ th feature,  $j = 1, 2, \dots, p$
- $x_i$  the  $i$ th instance,  $i = 1, 2, \dots, n$
- $\mu_{kj}$  the expectation of  $j$ th feature in  $C_k$ ,  $k = 1, 2$ ;  $j = 1, 2, \dots, p$
- $\sigma_{kj}^2$  the variance of  $j$ th feature in  $C_k$ ,  $k = 1, 2$ ;  $j = 1, 2, \dots, p$
- $\bar{x}_{kj}$  the sample mean of  $j$ th feature in  $C_k$ ,  $k = 1, 2$ ;  $j = 1, 2, \dots, p$
- $s_{kj}^2$  the sample variance of  $j$ th feature in  $C_k$ ,  $k = 1, 2$ ;  $j = 1, 2, \dots, p$

### 4.2. Eight Filtering Methods

#### 4.2.1. $t$ Test

Feature screening using the  $t$  test statistic [34] is similar to performing a hypothesis test (the null hypothesis is that there is no difference in the means) on the class's distribution, and its significance indicates the difference between majority and minority classes. The lower the  $p$  value of this  $t$  test, the higher the majority and minority classes' significant difference. Consequently, the considered feature is more relevant to the separation of two classes.



#### 4.2.2. Fisher Score

Fisher score [35] is simple and generally quite effective, which can be a criterion of feature screening. Fisher score of a single feature is defined as follows:

$$Fisher(x_j) = \frac{|\mu_{1j} - \mu_{2j}|}{\sigma_{1j}^2 + \sigma_{2j}^2}, \quad j = 1, 2, \dots, p, \quad (1)$$

$\mu_{1j}$ ,  $\mu_{2j}$ ,  $\sigma_{1j}^2$ , and  $\sigma_{2j}^2$  can be replaced by their corresponding sample statistics in computation, namely,

$$Fisher(x_j) = \frac{|\bar{x}_{1j} - \bar{x}_{2j}|}{s_{1j}^2 + s_{2j}^2}, \quad j = 1, 2, \dots, p \quad (2)$$

A feature with a large Fisher score is more crucial for discriminating the two categories.

#### 4.2.3. Hellinger Distance

Hellinger distance can be used to measure a distributional divergence [36]. Denoted the two normal distributions by  $P$  and  $Q$ , Hellinger distance is calculated as follows:

$$D_H^2(P, Q) = 2 - 2\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2} \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\}}, \quad (3)$$

where  $\mu_1$ ,  $\sigma_1^2$ ,  $\mu_2$ , and  $\sigma_2^2$  are the expectation and variance of  $P$  and  $Q$ , respectively, and their corresponding sample statistics are used in practice [37]. The larger the Hellinger distance is, the more divergent the two distributions are.

#### 4.2.4. Relief and ReliefF

The Relief is an iteration method that tries to give each feature a score to indicate its level of relevance to the response [37,38]. Let  $x_i$  be an instance;  $nearhit_i$  and  $nearmiss_i$  be its two nearest neighbors from the same class and the other class by the Euclidean distance, respectively. The score vector  $s = (s_1, s_2, \dots, s_p)^T$  is refreshed as follows:

$$s_j \leftarrow s_j - (x_{ij} - nearhit_{ij})^2 + (x_{ij} - nearmiss_{ij})^2, \quad j = 1, 2, \dots, p, \quad (4)$$

where  $x_{ij}$ ,  $nearhit_{ij}$  and  $nearmiss_{ij}$  are the  $j$ th element of  $x_i$ ,  $nearhit_i$  and  $nearmiss_i$ , respectively. A feature with a higher score is more crucial to the response. Though ReliefF [39] is originally developed for dealing with multi-class and noise datasets, it can be applied to binary classification cases. Compared with Relief that searches one nearest instance from the same class and one from the other class in updating the weights, ReliefF finds  $k$  nearest neighbors. Similarly, a feature with a higher score is more important to the response.

#### 4.2.5. Information Gain (IG)

Information gain [40] is the measurement of informational theory and can be utilized to assess the importance of a given feature. In the settings of binary classification, the information entropy of the set  $D$  is defined as follows:

$$Ent(D) = - \sum_{k=1}^2 \frac{n_k}{n} \log_2 \frac{n_k}{n} \quad (5)$$

Assuming that a discrete feature (attribute)  $x$  has  $V$  different values  $\{x_1, x_2, \dots, x_V\}$ , and  $D_v$  is the subset of instance set  $D$  satisfying  $x = x_v$ . The information gain of the variable  $x$  is

$$IG(D, x) = Ent(D) - \sum_{v=1}^V \frac{|D_v|}{n} Ent(D_v), \quad (6)$$

The larger the information gain is, the more important the feature is for separating the classes. A continuous feature should be discretized before using the IG metric.

#### 4.2.6. Gini Index

Gini index [41] fits binary digits, continuous numerical values, ordinal numbers, etc. It is a non-purity split method. Gini index of  $D$  is defined as follows:

$$Gini(D) = 1 - \sum_{k=1}^2 p_k^2, \quad (7)$$

where  $p_k$  is the probability that any instance belongs to  $C_k$ , and it is replaced with  $n_k/n$ , ( $k = 1, 2$ ) in practice. If we divide  $D$  into  $M$  subsets  $D_1, D_2, \dots, D_M$ , the Gini index after splitting is:

$$Gini_{split}(D) = \sum_{m=1}^M \frac{|D_m|}{n} Gini(D_m) \quad (8)$$

The smaller the Gini index is, the more important the feature is.

#### 4.2.7. R-Value

R-value [30,42] indicates the degree of overlap for the class-imbalanced dataset. R-value for a dataset  $D$  is defined as follows:

$$R(D) = \frac{1}{n} \sum_{k=1}^2 \sum_{m=1}^{|C_k|} \lambda(|kNN(P_{km}, D - C_k)| - \theta), \quad (9)$$

where

$$\lambda(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

where  $kNN(P, C_i)$  is the subset of  $k$  nearest neighbors of instance  $P$  that belong to the set of instances  $C_i$ , and  $\theta$  is the threshold generally set to be  $k/2$  [43]. The smaller the R-value is, the more important the feature is for discriminating the categories.

### 4.3. Four Evaluation Metrics

#### 4.3.1. Geometric Mean and F-Measure

True positive, true negative, false positive, and false negative are denoted by  $TP$ ,  $TN$ ,  $FP$ , and  $FN$ , respectively. Some common metrics are listed below:

$$\begin{aligned} TPR = recall &= \frac{TP}{TP + FN}; \\ TNR &= \frac{TN}{TN + FP}; \\ FPR &= \frac{FP}{FP + TN}; \\ precision &= \frac{TP}{TP + FP}; \\ Gmean &= \sqrt{TPR \times TNR}; \\ F_1 &= \frac{2precision \times recall}{precision + recall} \end{aligned}$$

The range of both  $Gmean$  and  $F_1$  is  $[0, 1]$ . The larger they are, the better the classifier works.

#### 4.3.2. AUCROC and AUCPRC

AUCROC is the area under the receiver operating characteristic curve (ROC) [44]. AUCPRC is the area under the precision recall curve (PRC) [45]. Both AUCROC and

*AUCPRC* range from 0 to 1, and the larger they are, the better the classifier is built for the imbalanced learning. More details on *AUCROC* or *AUCPRC* can be found in our previous studies [37,46].

*Gmean*,  $F_1$ , *AUCROC*, and *AUCPRC* are more widely used than the metric *Accuracy* in class-imbalance learning. These metrics actually pay more attention to the minority samples.

#### 4.4. Kendall's $\tau$ Rank Correlation

Kendall's  $\tau$  rank correlation statistic [47] can be applied to calculate the degree of comparability between the feature rankings of two filtering techniques. Let the two feature rankings generated by two filters be

$$\begin{aligned} f_1 &: r_{11}, r_{21}, r_{31}, \dots, r_{p1}, \\ f_2 &: r_{12}, r_{22}, r_{32}, \dots, r_{p2}, \end{aligned}$$

and there are no ties in each of ranking list. Then Kendall's  $\tau$  is calculated as follows,

$$\tau(f_1, f_2) = \frac{\sum_{i < j} \text{sgn}(r_{i1} - r_{j1}) \text{sgn}(r_{i2} - r_{j2})}{p(p-1)/2}, \quad (11)$$

where  $\text{sgn}(x)$  is the sign function, namely it equals 1 if  $x$  is positive and  $-1$  if  $x$  negative. A pair of  $(i, j)$  is called concordant if  $r_{i1} > r_{j1}$  and  $r_{i2} > r_{j2}$  or  $r_{i1} < r_{j1}$  and  $r_{i2} < r_{j2}$ . Otherwise, they are considered discordant. The numerator  $\sum_{i < j} \text{sgn}(r_{i1} - r_{j1}) \text{sgn}(r_{i2} - r_{j2})$

is the difference between the number of concordant pairs and the number of discordant pairs, and the denominator  $p(p-1)/2$  is the number of all distinct pairs of  $p$  elements. The range of  $\tau$  is  $[-1, 1]$ . If  $\tau = 0$ , the correlation of two rankings is weak; if  $\tau = -1$ , then all pairs will be discordant, and the two rankings are exactly opposite; if  $\tau = 1$ , then all pairs are exactly concordant [48].

#### 4.5. Rank Aggregation with Re-Balance for Class-Imbalanced Data

As mentioned above, there are differences among the ranks from different filtering methods, but we assume that they are equal in match, namely, no one is better or worse than another. Rank aggregation (RA) is a greatly intuitive metric that computes the absolute differences between the ranks of all individual features [49]. Rank aggregation with re-balance (RAR) consists of two stages for class-imbalanced data and is illustrated in Figure 11. In sample space, the data are artificially balanced by generating new instances of the minority class or (and) removing some of the majority class instances. In feature space,  $m$  rank lists are first computed using  $m$  different filtering methods. Each rank list is the full permutation of all the features. Then, they are merged to be an aggregated rank. Feature screening and classification can be performed according to this aggregated rank.

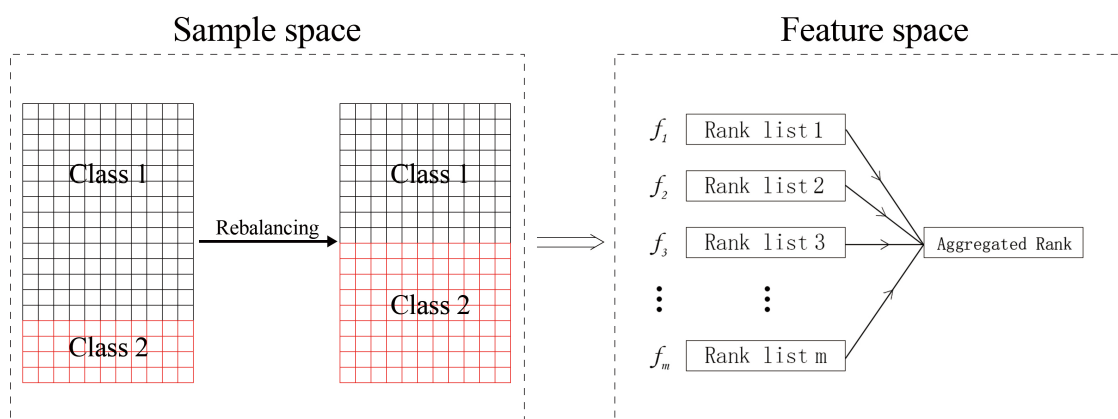


Figure 11. The frame of rank aggregation with re-balance.



#### 4.5.1. Rank Aggregation

As mentioned above, different filter techniques will give different feature ranking results. The rank aggregation method [34,50] combines all the rankings together, by aggregating all feature ranking lists generated from different filtering methods.

RA is to find an optimal ranking  $\delta^*$  such that

$$\delta^* = \arg \min \sum_{i=1}^K w_i d(\delta, f_i), \quad (12)$$

where  $f_i$  is the  $i$ th feature ranking list,  $\delta$  represents a ranking list with the same length of  $f_i$ ,  $d$  is a distance function, and  $w_i$  is the important weight related with list  $f_i$ . In this study,  $d$  is chosen to be the Spearman's foot rule distance [50]:

$$d(\delta, L_M) = \sum_{t \in L_M \cup \delta} |M(r^\delta(t)) - M(r^{L_M}(t))| \times |r^\delta(t) - r^{L_M}(t)| \quad (13)$$

$L_M = \{A_1^M, \dots, A_m^M\}$  denotes an ordered list of top  $m$  algorithms produced by the validation measure  $M$ . Let  $M(1), \dots, M(m)$  be the scores for the top  $m$  algorithms in  $L_M$ , where  $M(1)$  is the best score given by measure  $M$  and so on. Let  $r^M(A)$  be the rank of  $A$  under  $M$  (1 means "best") if  $A$  is within top  $m$ , and be equal to  $m + 1$ ; otherwise,  $r^\delta(A)$  is defined likewise.

The optimization of the objective (12) is achieved by using the Monte Carlo cross-entropy (CE) algorithm [51,52]. CE Monte Carlo algorithm is a stochastic search method, which produces a "better" sample in the future, which is concentrated around an  $x$  that corresponds to an optimal  $\delta^*$  [50].

#### 4.5.2. Strategies to Generate New Samples

Before performing rank aggregation, the training instances are to be modified to produce a more balanced class distribution. To achieve this task, new minority or (and) majority class samples need to be generated or drawn from the original dataset. We employ the following three strategies to gain new samples:

##### Randomly Sampling

In the over-sampling, some (all) the minority class instances are randomly duplicated; in the under-sampling, a portion of majority samples are randomly removed.

##### SMOTE

Synthetic minority over-sampling technique (SMOTE) is a popular over-sampling algorithm [5]. Figure 12 illustrates how to generate new samples according to the selected point  $\mathbf{x}_i$  in SMOTE. The five selected nearest neighbors of  $\mathbf{x}_i$  are  $\mathbf{x}_{i1}$  to  $\mathbf{x}_{i5}$ .  $\mathbf{x}'_{i1}$  to  $\mathbf{x}'_{i5}$  are the synthetic data points created by the randomized interpolation. Namely,

$$\mathbf{x}'_{ih} = \mathbf{x}_i + u_h(\mathbf{x}_{ih} - \mathbf{x}_i), \quad h = 1, 2, \dots, 5,$$

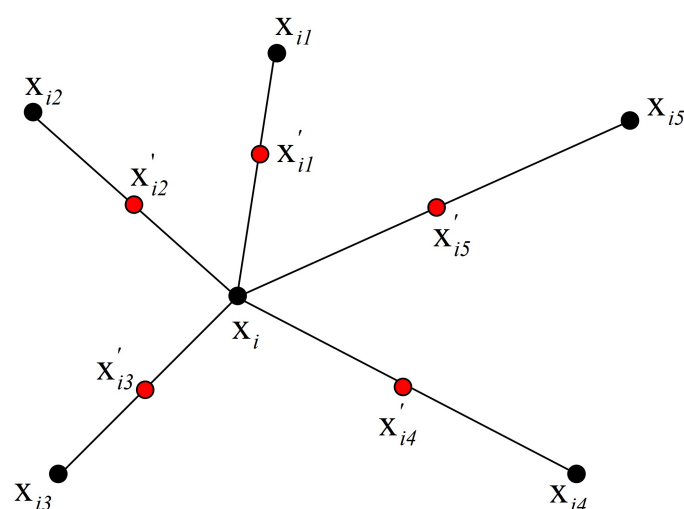
where  $u_h$  is a random number between 0 and 1. The above operation can be repeated to obtain requested synthetic minority instances.

##### Smoothed Bootstrap

Smoothed bootstrap technique repeatedly bootstraps the data from the two classes and employs smoothed kernel functions to generate new approximately balanced samples [53]. A new instance is generated by performing the following three steps:

- step 1: choose  $y = k \in \{1, 2\}$  with probability 0.5;
- step 2: choose  $(\mathbf{x}_i, y_i)$  in the original data set such that  $y_i = k$  with probability  $1/n_k$ ;
- step 3: sample  $\mathbf{x}$  from a probability distribution  $K_{\mathbf{H}_k}(\cdot, \mathbf{x}_i)$ , which is centered at  $\mathbf{x}_i$  and depends on the smoothing matrix  $\mathbf{H}_k$ .

In brief, smoothed bootstrap firstly draws randomly from the original dataset an instance from one of the two categories, then generates a new instance in its neighborhood.



**Figure 12.** An illustration of how to create the synthetic data points in the SMOTE algorithm.

#### 4.6. Experiment and Assessing Metrics

As shown in Table 5, five metabolomics datasets were employed to test our algorithm. NPC is a nasopharyngeal carcinoma dataset [32,54] that is exactly balanced. In this study, NPC was utilized to investigate the performance of rank aggregation strategy on original balanced data, which included 100 patients with nasopharyngeal carcinoma and 100 healthy controls. Traumatic brain injury (TBI) is from our previous studies [32,55], which reports the serum metabolic profiling of TBI patients with (or without) cognitive impairment (CI). The TBI dataset included 73 TBI patients with CI and 31 TBI patients without CI. CHD2-1 and CHD2-2 datasets are actually from the same experiment about coronary heart disease (CHD) [30]. The CHD2-1 dataset contains 21 patients with CHD, and the CHD2-2 dataset contains 16 patients with coronary heart disease associated with type 2 diabetes mellitus (CHD-T2DM), which are compared with a control group of 51 healthy adults. ATR is an Acori Tatarinowii Rhizoma dataset, which included 21 samples collected from Sichuan Province, and 8 samples were from Anhui Province in China [56]. Table 5 lists the summary of five datasets; included are the numbers of attributes, total instances, the majority, the minority instances, and the imbalance ratio. The NPC dataset was utilized to test the performance of rank aggregation under original balanced distribution. The other four imbalanced data sets were used to evaluate the RAR algorithm with artificially re-balanced data.

This section shows the efficacy of the proposed RAR algorithm on one original balanced dataset and four class-imbalanced datasets and compares it with other filtering feature screening methods via several assessing metrics. Rank aggregation was performed under the following seven situations:

- **Case 1:** no re-sampling: the original datasets are directly utilized to perform rank aggregation. Denoted case 1 by “RA” because there is no re-sampling in it.
- **Case 2:** hybrid-sampling A: some instances of the majority class are randomly eliminated, and new synthetic minority examples are generated by SMOTE. The size of the remaining majority is equal to the size of the (original plus new generated) minority class.
- **Case 3:** hybrid-sampling B: A new synthetic dataset is generated according to the smoothed bootstrap re-sampling technique. The sizes of the majority and minority classes are approximately equal.
- **Case 4:** over-sampling A: new minority class instances are randomly duplicated according to the original minority group.

- **Case 5:** over-sampling B: new synthetic minority examples are generated on the basis of the smoothed bootstrap re-sampling technique.
- **Case 6:** under-sampling A: some instances from the majority class are randomly removed so that the size of the remaining majority class is equal to the size of the minority.
- **Case 7:** under-sampling B: new synthetic majority examples are generated according to the smoothed bootstrap re-sampling technique.

Note that NPC is balanced, and just case 1 is performed on it. Table 6 lists the summary of the six re-balanced strategies. In this study,  $Gmean$ ,  $F_1$ ,  $AUCROC$ , and  $AUCPRC$  are employed to assess the performance of RA or RAR algorithm on five datasets under seven cases.

**Table 6.** Re-balanced strategies.

Methods	Re-Sampling Process			Algorithm Process		
	Under-Sampling	Over-Sampling	Hybrid	SMOTE	Random	Smoothed Bootstrap
Case 2			Yes	Yes		
Case 3			Yes			Yes
Case 4		Yes			Yes	
Case 5		Yes				Yes
Case 6	Yes				Yes	
Case 7	Yes					Yes

## 5. Conclusions

In this paper, we propose a simple but effective strategy called RAR for feature screening of class-imbalanced data by aggregating rankings from individual filtering algorithms and modifying the class-imbalanced data with various re-sampling methods to provide balanced or more adequate data. RAR can address the problem of inconsistency between different feature ranking methods to a large extent. The results on real datasets show that RAR is highly competitive and almost better than single filtering screening in terms of geometric mean, F-measure,  $AUCROC$ , and  $AUCPRC$ . After performing re-balanced pretreatment, the performance of rank aggregation can be highly improved, so re-sampling to balance the classes is extremely useful in rank aggregation when the data are class-imbalanced in metabolomics. Our proposed method serves as a reference for future research on feature selection for the diagnosis of diseases.

Rank aggregation is a general idea to investigate the importance of features. In this study, rankings from eight filtering algorithms are employed to generate the aggregated rank. There are many other filter techniques, such as Chi-squared, power, Kolmogorov–Smirnov statistic, and signal-to-noise ratio [57], which are all widely utilized in class-imbalance learning. In addition, considering that a re-sampling method can also generate a rank list, rank aggregation can be performed according to the various re-sampling algorithms rather than different filtering methods. Further, if necessary, ensemble multiple rank aggregations could be performed to combine those aggregated rankings derived from different algorithms. Finally, although RAR is used in the metabolomics datasets in this study, it is potentially available for handling high-dimensional imbalanced data from other fields, such as economics and biology.

**Author Contributions:** Conceptualization, G.-H.F.; Data curation, G.-H.F. and J.-B.W.; Funding acquisition, L.-Z.Y.; Software, J.-B.W.; Writing—original draft, M.-J.Z.; Writing—review and editing, G.-H.F. and L.-Z.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is financially supported by the National Natural Science Foundation of China (Grant Nos. 11761041 and 21775058).

**Institutional Review Board Statement:** Not applicable.



**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets for this article are available from the corresponding author.

**Acknowledgments:** The authors sincerely thank the academic editor and four anonymous reviewers for their constructive comments that led to the current improved version of the paper. We would like to thank the National Natural Science Foundation of China for its financial support (Grant Nos. 11761041 and 21775058).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Brodley, C.; Friedl, M. Identifying mislabeled training data. *J. Artif. Intell. Res.* **1999**, *11*, 131–167. [\[CrossRef\]](#)
2. Chawla, N. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 875–886.
3. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [\[CrossRef\]](#)
4. Cordon, I.; Garcia, S.; Fernandez, A.; Herrera, F. Imbalance: Oversampling algorithms for imbalanced classification in R. *Knowl. Based Syst.* **2018**, *161*, 329–341. [\[CrossRef\]](#)
5. Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, W. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
6. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *R J.* **2014**, *6*, 79–89. [\[CrossRef\]](#)
7. Hulse, J.V.; Khoshgoftaar, T.; Napolitano, A.; Wald, R. Feature selection with high-dimensional imbalanced data. In Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, Miami, FL, USA, 6 December 2009; pp. 507–514.
8. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
9. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [\[CrossRef\]](#)
10. Yun, Y.H.; Li, H.D.; Deng, B.C.; Cao, D.S. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends Anal. Chem.* **2019**, *113*, 102–115. [\[CrossRef\]](#)
11. Su, X.; Khoshgoftaar, T. A survey of collaborative filtering techniques. *Adv. Artif. Intell.* **2009**, *2009*, 421425. [\[CrossRef\]](#)
12. Ambjørn, J.; Janik, R.; Kristjansen, C. Wrapping interactions and a new source of corrections to the spin-chain/string duality. *Nucl. Phys. B* **2006**, *736*, 288–301. [\[CrossRef\]](#)
13. Higman, G.; Neumann, B.; Neuman, H. Embedding theorems for groups. *J. Lond. Math. Soc.* **1949**, *1*, 247–254. [\[CrossRef\]](#)
14. Guo, H.; Li, Y.; Shang, J.; Gu, M.; Huang, Y.; Gong, B. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239.
15. Gu, Q.; Li, Z.; Han, J. Generalized fisher score for feature selection. *arXiv* **2012**, arXiv:1202.3725.
16. Yin, L.; Ge, Y.; Xiao, K.; Wang, X.; Quan, X. Feature selection for high-dimensional imbalanced data. *Neurocomputing* **2013**, *105*, 3–11. [\[CrossRef\]](#)
17. Spolaôr, N.; Cherman, E.; Monard, M.; Lee, H. ReliefF for multi-label feature selection. In Proceedings of the 2013 Brazilian Conference on Intelligent Systems, Fortaleza, Brazil, 19–24 October 2013; pp. 6–11.
18. Kira, K.; Rendell, L. The feature selection problem: Traditional methods and a new algorithm. *Aaii* **1992**, *2*, 129–134.
19. Lee, C.; Lee, G. Information gain and divergence-based feature selection for machine learning-based text categorization. *Inf. Process. Manag.* **2006**, *42*, 155–165. [\[CrossRef\]](#)
20. Lerman, R.; Yitzhaki, S. A note on the calculation and interpretation of the Gini index. *Econ. Lett.* **1984**, *15*, 363–368. [\[CrossRef\]](#)
21. Lobo, J.; Jiménez-Valverde, A.; Real, R. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 145–151. [\[CrossRef\]](#)
22. Boyd, K.; Eng, K.; Page, C. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 451–466.
23. Altidor, W.; Khoshgoftaar, T.; Napolitano, A. Wrapper-based feature ranking for software engineering metrics. In Proceedings of the 2009 International Conference on Machine Learning and Applications, Miami, FL, USA, 13–15 December 2009; pp. 241–246.
24. Pillai, I.; Fumera, G.; Roli, F. F-measure optimisation in multi-label classifiers. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 2424–2427.
25. Lee, J.; Batnyam, N.; Oh, S. RFS: Efficient feature selection method based on R-value. *Comput. Biol. Med.* **2013**, *43*, 91–99. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Ali, M.; Ali, S.I.; Kim, D.; Hur, T.; Bang, J.; Lee, S.; Kang, B.H.; Hussain, M.; Zhou, F. UEFS: An efficient and comprehensive ensemble-based feature selection methodology to select informative features. *PLoS ONE* **2018**, *13*, e0202705. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Hoque, N.; Singh, M.; Bhattacharyya, D.K. EFS-MI: An ensemble feature selection method for classification. *Complex Intell. Syst.* **2018**, *4*, 105–118. [\[CrossRef\]](#)
28. Yang, P.; Liu, W.; Zhou, B.B.; Chawla, S.; Zomaya, A.Y. *Ensemble-Based Wrapper Methods for Feature Selection and Class Imbalance Learning*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 544–555.

29. Lin, X.; Yang, F.; Zhou, L.; Yin, P.; Kong, H.; Xing, W.; Lu, X.; Jia, L.; Wang, Q.; Xu, G. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J. Chromatogr. B* **2012**, *910*, 149–155. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Fu, G.H.; Wu, Y.J.; Zong, M.J.; Yi, L.Z. Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics. *Chemom. Intell. Lab. Syst.* **2020**, *196*, 103906. [\[CrossRef\]](#)
31. Sen, P. Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.* **1968**, *63*, 1379–1389. [\[CrossRef\]](#)
32. Fu, G.H.; Xu, F.; Zhang, B.Y.; Yi, L.Z. Stable variable selection of class-imbalanced data with precision-recall criterion. *Chemom. Intell. Lab. Syst.* **2017**, *171*, 241–250. [\[CrossRef\]](#)
33. Takaya, S.; Marc, R.; Guy, B. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432.
34. Yun, Y.H.; Deng, B.C.; Cao, D.S.; Wang, W.T.; Liang, Y.Z. Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery. *Anal. Chim. Acta* **2016**, *911*, 27–34. [\[CrossRef\]](#)
35. Weston, J.; Mukherjee, S.; Chapelle, O. Feature selection for SVMs. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 668–674.
36. Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [\[CrossRef\]](#)
37. Fu, G.H.; Wu, Y.J.; Zong, M.J.; Pan, J. Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data. *BMC Bioinform.* **2020**, *21*, 121. [\[CrossRef\]](#)
38. Robnik-Šikonja, M.; Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [\[CrossRef\]](#)
39. Kononenko, I. Estimating attributes: analysis and extensions of RELIEF. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 171–182.
40. Yang, Y.; Pedersen, J. A comparative study on feature selection in text categorization. *Icml* **1997**, *97*, 35.
41. Shang, W.; Huang, H.; Zhu, H.; Lin, Y.; Qu, Y.; Wang, Z. A novel feature selection algorithm for text categorization. *Expert Syst. Appl.* **2007**, *33*, 1–5. [\[CrossRef\]](#)
42. Borsos, Z.; Lemnaru, C.; Potolea, R. Dealing with overlap and imbalance: a new metric and approach. *Pattern Anal. Appl.* **2018**, *21*, 381–395. [\[CrossRef\]](#)
43. Oh, S. A new dataset evaluation method based on category overlap. *Comput. Biol. Med.* **2011**, *41*, 115–122. [\[CrossRef\]](#)
44. Provost, F.; Fawcett, T. Robust classification for imprecise environments. *Mach. Learn.* **2001**, *42*, 203–231. [\[CrossRef\]](#)
45. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240.
46. Fu, G.H.; Yi, L.Z.; Pan, J. Tuning model parameters in class-imbalanced learning with precision-recall curve. *Biom. J.* **2019**, *61*, 652–664. [\[CrossRef\]](#)
47. Kendall, M.G. A New Measure of Rank Correlation. *Biometrika* **1938**, *30*, 81–93. [\[CrossRef\]](#)
48. Shieh, G. A weighted Kendall's tau statistic. *Stat. Probab. Lett.* **1998**, *39*, 17–24. [\[CrossRef\]](#)
49. Pihur, V. *Statistical Methods for High-Dimensional Genomics Data Analysis*; University of Louisville: Louisville, KY, USA, 2009.
50. Pihur, V.; Datta, S.; Datta, S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinform.* **2009**, *10*, 62. [\[CrossRef\]](#)
51. Pihur, V.; Datta, S.; Datta, S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics* **2007**, *23*, 1607–1615. [\[CrossRef\]](#)
52. Pihur, V.; Datta, S.; Datta, S. Finding common genes in multiple cancer types through meta-analysis of microarray experiments: A rank aggregation approach. *Genomics* **2008**, *92*, 400–403. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Menardi, G.; Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Discov.* **2014**, *28*, 92–122. [\[CrossRef\]](#)
54. Fu, G.H.; Yi, L.Z.; Pan, J. LASSO-based false-positive selection for class-imbalanced data in metabolomics. *J. Chemom.* **2019**, *33*. [\[CrossRef\]](#)
55. Fu, G.H.; Zhang, B.Y.; Kou, H.D.; Yi, L.Z. Stable biomarker screening and classification by subsampling-based sparse regularization coupled with support vector machines in metabolomics. *Chemom. Intell. Lab. Syst.* **2017**, *160*, 22–31. [\[CrossRef\]](#)
56. Ma, S.S.; Zhang, B.Y.; Chen, L.; Zhang, X.J.; Ren, D.B.; Yi, L.Z. Discrimination of *Acori Tatarinowii* Rhizoma from two habitats based on GC-MS fingerprinting and LASSO-PLS-DA. *J. Cent. South Univ.* **2018**, *25*, 1063–1075. [\[CrossRef\]](#)
57. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. Dimensionality Reduction for Imbalanced Learning. In *Learning from Imbalanced Data Sets*; Springer: Cham, Switzerland, 2018; pp. 227–251. [\[CrossRef\]](#)