J. Meas. Sci. Instrum., 2025, 16(4): 612-624.
DOI: 10.62756/jmsi.1674-8042.2025059
http://jmsi.publish.founderss.cn
jmsi@nuc.edu.cn

Adaptive feature selection method for high-dimensional imbalanced data classification

WU Jianzhen[1], XUE Zhen[1,2*], ZHANG Liangliang[1], YANG Xu[1]

1. School of Mathematics, North University of China, Taiyuan 030051, China;
2. Department of Mathematics, City University of Hong Kong, Kowloon 999077, Hong Kong, China

***Corresponding author:** XUE Zhen (xuezhen@nuc.edu.cn)

Received: Apri 14, 2025 **Revised:** May 30, 2025 **Accepted:** July 9, 2025

Abstract: Data collected in fields such as cybersecurity and biomedicine often encounter high dimensionality and class imbalance. To address the problem of low classification accuracy for minority class samples arising from numerous irrelevant and redundant features in high-dimensional imbalanced data, we proposed a novel feature selection method named AMF-SGSK based on adaptive multi-filter and subspace-based gaining sharing knowledge. Firstly, the balanced dataset was obtained by random under-sampling. Secondly, combining the feature importance score with the AUC score for each filter method, we proposed a concept called feature hardness to judge the importance of feature, which could adaptively select the essential features. Finally, the optimal feature subset was obtained by gaining sharing knowledge in multiple subspaces. This approach effectively achieved dimensionality reduction for high-dimensional imbalanced data. The experiment results on 30 benchmark imbalanced datasets showed that AMF-SGSK performed better than other eight commonly used algorithms including BGWO and IG-SSO in terms of $F1$-score, AUC, and G-mean. The mean values of $F1$-score, AUC, and G-mean for AMF-SGSK are 0.950, 0.967, and 0.965, respectively, achieving the highest among all algorithms. And the mean value of G-mean is higher than those of IG-PSO, ReliefF-GWO, and BGOA by 3.72%, 11.12%, and 20.06%, respectively. Furthermore, the selected feature ratio is below 0.01 across the selected ten datasets, further demonstrating the proposed method's overall superiority over competing approaches. AMF-SGSK could adaptively remove irrelevant and redundant features and effectively improve the classification accuracy of high-dimensional imbalanced data, providing scientific and technological references for practical applications.

Key words: high-dimensional imbalanced data; adaptive feature selection; adaptive multi-filter; feature hardness; gaining sharing knowledge based algorithm; metaheuristic algorithm

0 Introduction

In the cybersecurity field, wireless sensor networks (WSNs) are composed of numerous sensor nodes deployed in monitoring regions for collecting data and transmitting it for analysis, enabling critical applications in military and industrial fields[1,2]. However, without adequate security measures, WSNs are vulnerable to cyber-attacks, where machine learning based intrusion detection systems (IDS) can effectively prevent attacks and threats to WSNs and achieve high accuracy[3]. However, in most cybersecurity datasets, normal data constitutes the vast majority, while attack data accounts for only a small proportion. Additionally, these datasets are characterized by a high number of features, among which some are irrelevant to classification. It not only increases the complexity of data processing, but also affects the accuracy of classification tasks. In the biomedical field, mass spectrometry, DNA microarray techniques generate a large amount of data, and machine learning can efficiently extract useful knowledge from these data, with its rapidity and high accuracy significantly reducing workloads[4]. But in medical datasets, patient samples are far fewer than normal samples, and the data dimensionality may reach thousands or even tens of thousands. It can be seen that the data collected in the above fields shares a common characteristic: the challenges of high dimensionality and class imbalance, which poses significant challenges to traditional machine learning models. For imbalanced data, traditional classification algorithms tend to be biased toward the majority class due to optimizing the overall classification accuracy, resulting in poor classification performance of the minority class. The data is usually sparse due to excessive features, making samples appear as outliers in high-dimensional imbalanced data,

leading to the weakening of the distinction between the majority class and the minority class, thus exacerbating the imbalanced problems. Moreover, numerous irrelevant and redundant features in the data lead to degraded classification performance and computational efficiency[5,6].

Generally speaking, there are two approaches to solving the problem of imbalanced data classification: algorithm-level methods and data-level methods[7]. Algorithm-level methods include techniques such as cost-sensitive learning and ensemble learning[8,9]. Data-level methods are widely used due to their convenience and can be categorized into under-sampling, oversampling, and hybrid sampling which fuses under-sampling and oversampling. Under-sampling method balances the dataset by reducing the majority class samples, which may result in the loss of information. Oversampling is more prevalent for low-dimensional data. Chawla et al.[10] proposed synthetic minority over-sampling TEchnique (SMOTE) which is the most widely used oversampling method. However, SMOTE fails to consider the differences among the minority class samples, potentially generating noisy and overlapping samples. Therefore, researchers have proposed various variant approaches to SMOTE. Borderline-SMOTE primarily focuses on oversampling minority class samples in boundary regions[11]. ADASYN assigns different weights to minority class samples and generates a varying number of synthetic samples for each minority class according to these weights[12]. SMOTE-RSB* improves sampling quality by incorporating rough set theory to remove low-quality synthetic samples during the sampling process[13]. MWMOTE first identifies hard-to-learn minority class samples, then assigns weights to these samples, and finally synthesizes samples using clustering-based methods[14].

When dealing with high-dimensional imbalanced data, the samples are directly synthesized in the high-dimensional space, tending to generate samples that are more likely to be outliers and fail to preserve the feature correlations in the original space. The usual processing method is to reduce dimensionality through feature extraction or feature selection, because feature extraction usually causes the loss of useful information, while useless features are mapped to mask the useful feature information. This not only disrupts the relationships between the original features but also constructs a new space that is less relevant to the identification of the minority class samples[15]. Whereas feature selection retains the original representation of the data features with superior readability and interpretability. Consequently, there are many combined feature selection methods to deal with high dimensional imbalanced data[16,17]. However, most of the methods directly perform feature selection for the data without considering the influence of imbalanced factors. This makes it easy to select the features that are unfavorable to the classification of the minority class samples, which inadequately removes irrelevant and redundant features in the data, and limits the data classification accuracy improvement.

To address the challenge posed by numerous irrelevant and redundant features in high-dimensional imbalanced data, which leads to low data classification accuracy, we proposed a feature selection method based on adaptive multi-filter and subspace-based gaining sharing knowledge. The method first used random under-sampling to reduce the effect of data imbalance on feature selection, then combined the feature importance scores from multi-filter methods and further proposed a concept termed "feature hardness", which can adaptively select the important features to eliminate irrelevant and redundant features. The optimal feature subset is ultimately obtained by gaining sharing knowledge-based algorithm performed in multiple subspaces with different functions. The main contributions of this paper to high-dimensional imbalanced data classification are summarized as follows.

1) We proposed an adaptive feature selection method named AMF-SGSK for high-dimensional imbalanced data, which combined two-stage feature selection, adaptive multi-filter, and subspace-based gaining-sharing knowledge, effectively enhancing the classification accuracy of high-dimensional imbalanced data.

2) We proposed a novel metric, feature hardness, to quantitatively evaluate feature importance, adaptively selecting discriminative features while eliminating irrelevant and redundant ones.

3) Different forms of gaining-sharing knowledge in multiple subspaces enabled the proposed algorithm to better select the optimal feature subset, thereby boosting classification performance.

1 Related work

Feature selection is an effective technique for handling high-dimensional data, aiming to extract the most discriminative feature subset for classification tasks. The primary approaches include filter, wrapper, and embedded methods. Filter methods rank features according to statistical metrics, independent of the classifier, with the advantages of low time complexity and high scalability. The widely used filter methods include chi-square test (CS)[18], information gain (IG)[19], minimum redundancy maximum relevance (mRMR)[20], and ReliefF[21] among others.

Although the filter methods are computationally efficient, the selected features might not necessarily be the optimal subset for classification. Wrapper methods evaluate the performance of feature subsets using search algorithms and classifiers. Although they are computationally more expensive, they often yield more accurate results than filter methods. Among them, metaheuristic algorithms are widely used because they can mitigate local optima to some extent and outperform other wrapper methods in terms of classification accuracy[22]. Representative metaheuristic algorithms include genetic algorithm (GA)[23], particle swarm optimization (PSO)[24], ant colony optimization (ACO)[25], grey wolf optimizer (GWO)[26], grasshopper optimization algorithm (GOA)[27] and so on. To overcome the drawbacks of filter and wrapper methods, hybrid approaches combining both have emerged as effective solutions for high-dimensional data. They are commonly adopted methods to enhance the computational efficiency of wrapper methods, which can still maintain high accuracy, and their effectiveness has been demonstrated[28-30].

Feature selection can effectively remove irrelevant and redundant features, making it often be used to deal with high-dimensional imbalanced data. Xu et al.[31] proposed a novel ensemble approach based on multiview optimization. Multiple feature subviews were generated by applying weighted random forests on the original data, followed by neighborhood component analysis for each subview to optimize features, alleviating the impacts of redundant and invalid features. Zhang et al.[32] improved SVM-RFE by replicating misclassified boundary minority class samples to enhance the focus on the minority class during feature selection and then performing SMOTE on the feature subsets of boundary minority class samples, improving the classification accuracy for high-dimensional imbalanced data. Aydogan et al.[33] proposed a classification method CBR-PSO based on rough set theory and particle swarm optimization algorithm for high-dimensional imbalanced data, which consists of attribute reduction and classification, effectively eliminating irrelevant and redundant features. Zhang et al.[34] proposed a particle swarm optimization-based fuzzy clustering feature selection method (PSOFS-FC), which can be more advantageous for handling high-dimensional imbalanced data with missing values at significantly reduced runtime. Almotairi et al.[35] proposed a feature selection method based on evolutionary population dynamics (EPD) and marine predator algorithm (MPA), which addresses the issues of premature convergence and insufficient search capability of MPA and has a superior effect for high-dimensional imbalanced biomedical data classification.

For hybrid feature selection methods of dealing with high-dimensional imbalanced data, Moayedikia et al.[36] proposed a feature selection method, SYMON, which combines symmetric uncertainty with harmony search, and experimental results yield better performance than other state-of-the-art methods. Abdulrauf et al.[37] first filtered features by combining multiple filter methods to overcome the disadvantage of biased results of individual filter methods under certain datasets, then removed redundant features by correlation-based redundancy methods, and finally used binary grasshopper optimization algorithm to obtain the optimal feature subsets. Sharifai et al.[38] proposed a hybrid approach by using six filter methods and then combining simulated annealing with grasshopper optimization algorithm, which yields better results through the combination of two metaheuristics. Sahu et al.[39] assembled the top-ranked features of information gain, mRMR, and ReliefF, then removed the duplicates as feature subsets, and finally used political optimizer to obtain the final feature subsets, achieving better results on high-dimensional multi-class imbalanced data. Liu et al.[40] proposed the GU-MOACOFS method, which combines symmetric uncertainty, genetic algorithm-based under-sampling, multi-objective ant colony optimization algorithm with ensemble learning, and shows superiority in both high and low-dimensional imbalanced data. Kim et al.[41] designed ELFS, a hybrid feature selection framework based on ensemble learning, which consists of three modules, namely preprocessing, representative classifiers selection, and feature selection based on conditional mutual information maximization, to efficiently find the optimal feature subset and deal with imbalanced data robustly.

The aforementioned hybrid feature selection methods for high-dimensional imbalanced data can achieve better results, however, there are still some drawbacks. 1) Using filter method to directly select features for high-dimensional imbalanced data without considering the effect of imbalanced factors tends to select features that are unfavorable to the minority class samples classification during filtering. 2) Treating all filter methods and features equally in multi-filter feature selection may affect the overall classification accuracy by selecting irrelevant features and filtering out useful features. 3) In metaheuristic feature selection, the features selected for binary conversion are easily affected by the S-shaped and V-shaped functions, and the difference between the results obtained by different algorithms with different functions may be large. To solve the above problems, we proposed a feature selection method based on adaptive multi-filter and subspace-based

gaining sharing knowledge.

2 Adaptive feature selection method

2.1 Adaptive multi-filter feature

Multi-filter can effectively alleviate the problem of biased results of a single filter on certain datasets, but if the filters are treated equally without considering the weights. When certain filters do not perform well on a particular dataset, some of the features selected by them will affect the subsequent selection, which may reduce the overall classification accuracy. Similarly, uniform treatment of all features may lead to some useless features being selected and important features being filtered. Additionally, when applying filter methods to deal with high-dimensional imbalanced data, the impact of class imbalance also needs to be taken into account, which would otherwise result in the selection of features that are more biased towards the majority class.

By considering the importance of each feature with the classification effect of the filter method on the data, adaptive multi-filter (AMF) is proposed to remove irrelevant and redundant features from high-dimensional imbalanced data. In this paper, we employed six filter-based feature selection methods, namely CS, analysis of variance (ANOVA)[42], IG, Pearson correlation coefficient (PCC)[43], ReliefF and fast correlation-based filter (FCBF)[44]. To avoid the effect of class imbalance, the original dataset is first balanced using random under-sampling which randomly selects majority samples to remove. Subsequently, the six methods are applied to the balanced dataset, and each method selects the top 10% of features as the feature subset based on the its feature importance scores. The resulting six subsets are then classified using the KNN classifier to obtain their respective AUC scores. Different filter methods yield varying performance results on the same dataset. To avoid the influence of the poorly performing subsets, we use the AUC scores as the criterion to select the filter methods with the top n AUC scores, where n represents the number of selected methods. When n is too small, important features identified by other high-performance filter methods may be overlooked. Conversely, when n is too large, poor-performance filter methods may select more useless features, affecting the final selection results. In this paper, n is set to 3.

Among the first n selected filters, we first utilized the feature importance scores from each filter, and then employed the AUC scores of the subset as weights to combine the two. By introducing the concept of feature hardness in the form of Eq. (1) to represent the feature importance, we can adaptively select superior features based on their magnitude. To avoid different interval ranges of feature importance scores of different filter methods, the feature importance scores of selected n filter methods need to be normalized so that the score range is compressed to $[0,1]$.

$$FH_j=\frac{\sum_{i=1}^{n} auc_i \cdot s_{ij}^{\text{norm}}}{\sum_{i=1}^{n} auc_i}, \tag{1}$$

where FH_j denotes the feature hardness magnitude of the j-th feature, $s_{ij}^{\text{norm}}=\dfrac{s_{ij}-s_i^{\min}}{s_i^{\max}-s_i^{\min}}$ denotes the normalized score, s_{ij} denotes the score of the j-th feature in the i-th filter. $s_i^{\min}$ and $s_i^{\max}$ denote the minimum and maximum feature importance scores, respectively, for the j-th filter. auc_i denotes the AUC score of the subset of the i-th filter method. Based on the feature hardness magnitude, features with higher values indicate greater importance. The final feature subset is constructed by selecting features from the original dataset whose feature hardness scores exceed a predefined threshold. When the threshold is set too low, an excessive number of features are selected, potentially introducing useless features into the subset and complicating the optimization algorithm's search for the optimal feature combination. Conversely, if a threshold is set too high, it may filter out critical features containing valuable information, rendering it impossible to achieve the optimal feature subset regardless of the search strategy employed. The flowchart of the AMF is shown in Fig.1.

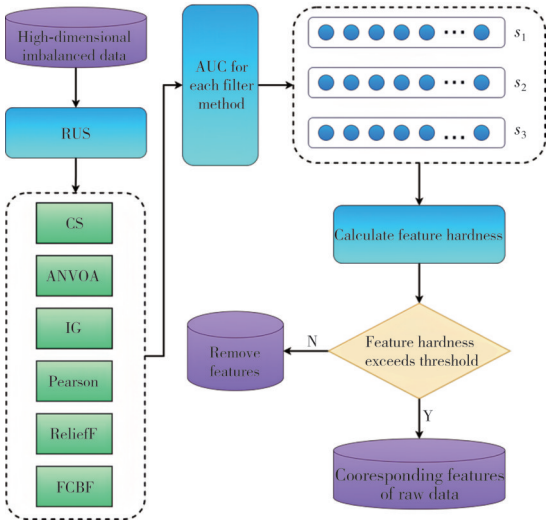

Fig. 1 Flowchart of adaptive multi-filter

Although the proposed AMF effectively eliminates numerous irrelevant and redundant features, it does not mean that the selected features are the optimal feature

subset for classification tasks. Nevertheless, AMF selects the high-quality features as much as possible, thereby facilitating subsequent metaheuristic feature selection to find the globally optimal feature subset.

2.2 Subspace-based gaining sharing knowledge-based algorithm

Gaining sharing knowledge-based algorithm (GSK) is a metaheuristic algorithm that progressively updates knowledge by modeling the gaining and sharing knowledge throughout the human lifetime[45]. The algorithm consists of the junior stage and the senior stage. Let the individuals of the population be represented as $\boldsymbol{x}_i$, $i=1,2,\cdots,N$, where N denotes the population size, $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{iD})$, D is the number of dimensions, and the number of dimensions D_{junior} and D_{senior} using the junior and senior stages are dynamically determined before each iteration according to

$$D_{\text{junior}}=D(1-G/G_{\max})^k, \tag{2}$$

$$D_{\text{senior}}=D-D_{\text{junior}}, \tag{3}$$

where k denotes the knowledge rate and is a positive real number, G is the current iteration number, and $G_{\max}$ is the maximum iteration number. As the number of iterations increases, the number of dimensions updated by the individual through the junior stage will decrease, and the senior stage will increase.

The steps of the junior stage of GSK are as follows. Individuals are ranked in descending order based on their fitness $f(\boldsymbol{x}_i)$, $i=1,2,\cdots,N$, generating an ordered sequence $(\boldsymbol{x}_{\text{best}},\cdots,\boldsymbol{x}_{i-1},\boldsymbol{x}_i,\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_{\text{worse}})$. For each individual $\boldsymbol{x}_i$, select the individuals $\boldsymbol{x}_{i-1}$ and $\boldsymbol{x}_{i+1}$ whose fitness values are adjacent to each other, and then randomly select another individual $\boldsymbol{x}_r$ as a knowledge-sharing source, and individual $\boldsymbol{x}_i$ is updated by

$$\boldsymbol{x}_i^{\text{new}}=\begin{cases}\boldsymbol{x}_i+k_f[(\boldsymbol{x}_{i-1}-\boldsymbol{x}_{i+1})+(\boldsymbol{x}_r-\boldsymbol{x}_i)], & f(\boldsymbol{x}_i)>f(\boldsymbol{x}_r),\\ \boldsymbol{x}_i+k_f[(\boldsymbol{x}_{i-1}-\boldsymbol{x}_{i+1})+(\boldsymbol{x}_i-\boldsymbol{x}_r)], & \text{else},\end{cases} \tag{4}$$

where the knowledge factor k_f is a real number greater than 0 that controls the amount of knowledge gained and shared from other individuals.

The steps of the senior stage of GSK are as follows. Based on individual fitness rankings, the population is stratified into three groups. The top $100p\%$ of individuals are called the best group, the bottom $100p\%$ of individuals are called the worst group, and the remaining individuals are called the middle group, where p is a real number from 0 to 0.5. Each x_i is updated according to

$$\boldsymbol{x}_i^{\text{new}}=\begin{cases}\boldsymbol{x}_i+k_f[(\boldsymbol{x}_{\text{pbest}}-\boldsymbol{x}_{\text{pworst}})+(\boldsymbol{x}_{\text{pmid}}-\boldsymbol{x}_i)], & f(\boldsymbol{x}_i)>f(\boldsymbol{x}_r),\\ \boldsymbol{x}_i+k_f[(\boldsymbol{x}_{\text{pbest}}-\boldsymbol{x}_{\text{pworst}})+(\boldsymbol{x}_i-\boldsymbol{x}_{\text{pmid}})], & \text{else},\end{cases} \tag{5}$$

where $\boldsymbol{x}_{\text{pbest}}$, $\boldsymbol{x}_{\text{pworst}}$ and $\boldsymbol{x}_{\text{pmid}}$ denote random individuals in the best, worst, and middle groups, respectively.

To address the feature selection problem, binary optimization of population individuals is required. Different S-shape or V-shape functions make the subset performance vary widely for different datasets. To alleviate the above problem, the feature subset obtained from AMF is replicated into four subsets. The GSK algorithm then applies different S-shaped and V-shaped transfer functions in the 4 subspaces for feature selection, and finally the subset with the highest fitness value is selected as the optimal feature subset. The S-shaped and V-shaped transfer functions used in this study are

$$S_1(x)=\frac{1}{1+\exp(-x)}, \tag{6}$$

$$S_2(x)=\frac{1}{1+\exp(-x/3)}, \tag{7}$$

$$V_1(x)=\left|\tanh(x)\right|, \tag{8}$$

$$V_2(x)=\left|\frac{2}{\pi}\arctan\left(\frac{\pi x}{2}\right)\right|. \tag{9}$$

Comparing the value of S-shape or V-shape function with the random real number in the interval (0, 1) expressed by rand(•), construct a piecewise function, $B(x_{ij}^{\text{new}})$. Taking S_1 function as an example, the corresponding formula is

$$B(x_{ij}^{\text{new}})=\begin{cases}1, & S_1(x_{ij}^{\text{new}})\geqslant \text{rand}(\bullet),\\ 0, & \text{else}.\end{cases} \tag{10}$$

If $B(x_{ij}^{\text{new}})$ is equal to 1, then keep the corresponding feature, and if $B(x_{ij}^{\text{new}})$ is equal to 0, then remove the feature.

The adaptive function serves as the judgment criterion for the metaheuristic feature selection effect. The traditional judgment criteria fails to account for the impact of data imbalance, usually using the overall accuracy of the data and the degree of feature deletion, which ignores the classification performance of minority class samples, so this paper adopts the common evaluation metric AUC with imbalanced data as the adaptive function.

The population size is one of the important parameters in the optimization algorithm. To improve the performance of the algorithm and avoid the influence of poorer individuals, the population size is reduced by deleting the individuals with the worst fitness values during the iteration process, which speeds up the convergence at a later stage. The population size for each iteration is

$$N_G=\text{round}[(N_{\min}-N_{\max})\bullet(G/G_{\max})+N_{\max}], \tag{11}$$

where $N_{\min}$ and $N_{\max}$ represent the minimum and maximum population sizes which are set to be 40 and 50, respectively. The flowchart of AMF-SGSK is illustrated in Fig.2.

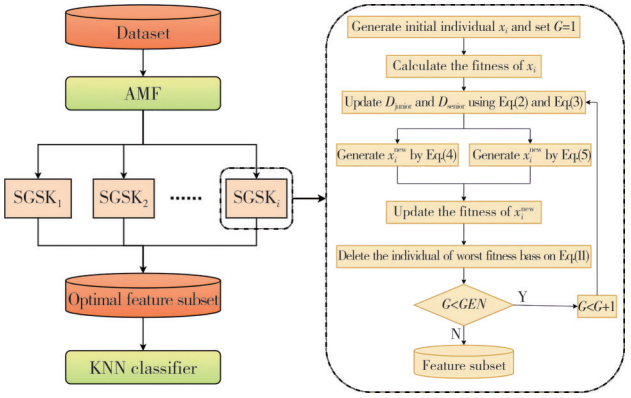

Fig. 2 Flowchart of AMF-SGSK

2.3 Time complexity and algorithm pseudocode

For high-dimensional imbalanced data, the time complexity of AMF-SGSK is calculated as follows. First, assuming the dataset has dimensionality m and sample size n, in the AMF phase, filter methods traverse n samples to statistically analyze each feature, resulting in a time complexity approximately $O(mn)$. In the SGSK phase, suppose the feature dimensionality is reduced to D after filtering, with iteration number G and population size N. And then the time complexity is approximately $O(DGN)$. Therefore, the overall time complexity of the algorithm is $O(mn+DGN)$. The pseudocode of AMF-SGSK is as follows.

Algorithm: AMF-SGSK
Input:
· X: An imbalanced dataset, f: filters, N: number of selected features, df: different functions
Output:
· X^{new}: the best subset
Begin:
1. **for** each filter $f_i(1\leqslant i\leqslant 6)$
2. obtain feature subset X_i'
3. calculate auc_i of f_i on X_i' by KNN classifier
4. **end for**
5. obtain optimal filters of by auc_i
6. **for** each optimal filter $of_i(1\leqslant i\leqslant 3)$
7. obtain s_{ij} of of_i on X
8. calculate s_{ij}^{norm} by Eq. (1)
9. **end for**
10. calculate FH_j by Eq. (2) $(1\leqslant j\leqslant N)$
11. obtain feature subset X_{AMF}
12. **for** k=1 to 4
13. apply GSK to X_{AMF} in different functions df_k
14. remove the worst individual during the iteration process
15. obtain the subset X_k^{new}
16. **end for**
17. select the best subset X^{new} with the highest fitness among the 4 subsets
end

3 Experimental results and analysis

KNN is selected as the classifier in the experiment. It offers stronger model interpretability and higher computational efficiency than random forest and SVM. Additionally, it is less prone to overfitting when the dataset size is small, making it a frequent choice as a classifier in feature selection research[46]. All optimization-based methods adopt the AUC metric under this classifier as the fitness function, with the maximum size of iterations set to 100 and the population size to 50. To provide reliable model performance estimation, all algorithms in the following experiments are evaluated on the dataset using 5-fold cross-validation, with the average value used as the final result.

3.1 Dataset

Thirty datasets from Open ML[47], UCI[48], and ASU[49] repositories were selected to validate the performance of the algorithm. The details of the datasets are summarized in Table 1, with feature counts ranging from 856 to 16 063 and most being from the biomedical domain.

Table 1 Detailed description of datasets

ID	Dataset	Instance	Features	IR	ID	Dataset	Instance	Features	IR
D1	Parkinsons disease classification	756	754	2.94	D16	leukemia	72	7 129	1.88
D2	QSAR androgen receptor	1 687	1 024	7.48	D17	tumors_C	60	7 129	1.86
D3	Period changer	90	1 177	2.33	D18	nci9-1	60	9 712	5.67
D4	colon	62	2 000	1.82	D19	nci9-2	60	9 712	5.67
D5	lung2	203	3 312	10.94	D20	amazon-commerce-reviews-3_vs_1-2-4-5	150	10 000	4.00
D6	lung3	203	3 312	8.67	D21	amazon-commerce-reviews-4_vs_1-2-4-5	150	10 000	4.00
D7	Olivetti_Faces0	400	4 096	39.00	D22	Umistfacescropped0	575	10 304	14.13
D8	Olivetti_Faces2	400	4 096	39.00	D23	AP_Prostate_Kidney	329	10 936	3.77
D9	GLIOMA1	50	4 434	2.57	D24	CLL_SUB_111-2	111	11 340	1.27
D10	GLIOMA3	50	4 434	2.57	D25	11_Tumors0	174	12 533	5.44
D11	DBWorld e-mails	64	4 702	1.21	D26	11_Tumors2	174	12 533	5.69
D12	TOX_171-3	171	5 748	3.38	D27	MLL1	72	12 582	2.00
D13	eating_0vs_123	553	6 373	2.95	D28	MLL2	72	12 582	1.57
D14	eating_1vs_023	553	6 373	2.95	D29	Ovarian	253	15 154	1.78
D15	eating_2vs_013	553	6 373	3.16	D30	GCM3	190	16 063	8.50

3.2 Evaluation metrics and data preprocessing

The evaluation metrics of classification performance for imbalanced data can be derived from the confusion matrix, as shown in Table 2, which contains TP (true positive), FN (false negative), FP (false positive) and

TN (true negative).

Table 2 Confusion matrix

Actual results	Prediction results	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Precision, recall, specificity, $F1$-score, G-mean, and AUC[50] can be calculated from the confusion matrix.

1) Precision is defined as the proportion of correct predictions among all instances predicted as the minority class.

$$Precision = \frac{TP}{TP + FP}. \tag{12}$$

2) Recall is defined as the proportion of correctly predicted positive instances to actual positive instances.

$$Recall = \frac{TP}{TP + FN}. \tag{13}$$

3) Specificity is defined as the proportion of correctly predicted negative instances to actual negative instances.

$$Specificity = \frac{TN}{TN + FP}. \tag{14}$$

4) $F1$-score is defined as the harmonic mean of precision and recall.

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{15}$$

5) G-mean is the geometric mean of recall and specificity

$$G\text{-}mean = \sqrt{Recall \cdot Specificity}. \tag{16}$$

6) AUC is the area under the ROC curve which is plotted with TPR as the horizontal axis and FPR as the vertical axis. The corresponding formulas are

$$TPR = \frac{TP}{TP + FN}, \tag{17}$$

$$FPR = \frac{FP}{TN + FP}, \tag{18}$$

$$AUC = \frac{1 + TPR + FPR}{2}. \tag{19}$$

In this study, we employed $F1$-score, AUC, and G-mean as the evaluation metrics of classification performance. Before conducting the experiments, data preprocessing is required. To avoid issues such as slow model training and insignificant accuracy improvement caused by large magnitude gaps between data dimensions, min-max normalization is applied to map the data to the range of $(0, 1)$. The datasets used in this study contain no missing values. If any missing values are present in a small number of samples, those samples are directly removed.

3.3 Threshold setting of feature hardness

Feature hardness threshold is an important parameter in adaptive multi-filter feature selection, significantly influencing the search performance of the optimization algorithm. When the threshold is set too high, it may result in fewer features being retained, potentially deleting some important feature information and making it impossible to obtain the optimal feature subset from the remaining features. Conversely, if the threshold is set too low, it may lead to an excessive number of retained features, which makes it difficult for SGSK to search for a selection scheme with the optimal feature subset. We selected six datasets with different dimensions that generalize the data dimensions used in Table 1. The comparative effect of $F1$-score of AMF-SGSK under different thresholds is shown in Fig. 3. To make the comparison more obvious, we normalized the $F1$-score of each dataset under different thresholds.

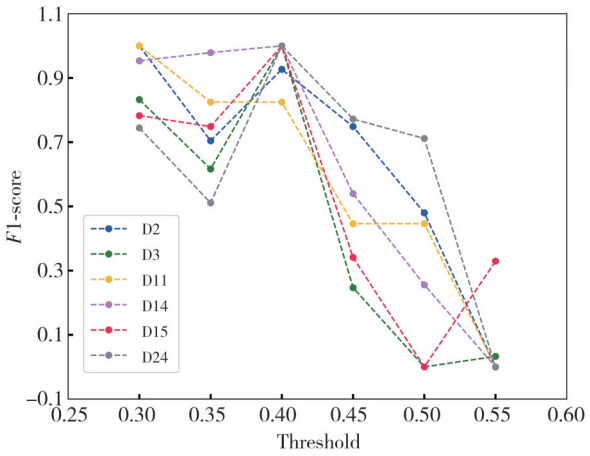

Fig. 3 Comparison of $F1$-score across different thresholds

As shown in Fig. 3, the $F1$-score values of D3, D14, and D24 datasets exhibit an increasing first and then decreasing trend as the threshold increases. For D11, the $F1$-score decreases consistently with increasing thresholds. Additionally, when the threshold is too small, an excessive number of features leads to a sharp increase in computational complexity during SGSK's subset search. To prevent excessively slow search speeds, when the classification performance gap is minimal across different thresholds, a higher threshold is preferred. Therefore, the feature hardness threshold is overall set within the range [0.33, 0.43], achieving better performance across all datasets.

3.4 Ablation experiment

To validate the effectiveness of each component, we conducted ablation studies on different models. Table 3

represents the mean values of *F*1-score, AUC, and *G*-mean for RAW, AMF, SGSK, and AMF-SGSK methods.

Table 3 Comparison of average results across different evaluation metrics

Model	AMF	SGSK	*F*1-score	AUC	G-mean
RAW			0.567 7	0.739 8	0.625 2
AMF	√		0.769 6	0.850 8	0.815 7
SGSK		√	0.793 0	0.868 5	0.816 8
AMF-SGSK	√	√	0.948 0	0.965 5	0.963 6

Table 3 demonstrates that all components of AMF-SGSK significantly outperform the RAW, with each module contributing to measurable performance improvements. The results of AMF are better than RAW, indicating that the method effectively removes irrelevant and redundant information and improves classification accuracy. AMF-SGSK achieves the best performance by selecting informative features by AMF and then optimizing them, avoiding the difficulty of obtaining the optimal selection scheme by directly performing SGSK in high-dimensional space.

3.5 Performance analysis of adaptive multi-filter

The results of AMF were compared with six filter methods (CS, ANOVA, IG, PCC, ReliefF, and FCBF) combined with random under-sampling. Table 4 presents the datasets on which AMF achieved its highest AUC performance, where bold indicates the maximum value for all methods.

Table 4 Comparison of AUC results across filter methods

ID	CS	ANOVA	IG	Pearson	ReliefF	FCBF	AMF
D1	0.767 7	0.747 4	0.735 7	0.747 4	0.605 5	0.771 1	**0.792 4**
D5	0.936 3	**0.972 3**	**0.972 3**	**0.972 3**	0.933 6	0.967 0	**0.972 3**
D7	0.850 0	0.894 9	0.847 4	0.894 9	0.750 0	0.800 0	**0.947 4**
D8	0.646 2	0.697 4	0.744 9	0.697 4	0.550 0	0.650 0	**1.000 0**
D10	0.773 2	0.736 9	0.749 4	0.736 9	0.758 9	0.838 1	**0.852 4**
D11	0.527 1	0.597 1	0.703 8	0.583 8	0.539 0	0.580 5	**0.820 5**
D13	0.574 0	0.570 4	0.533 6	0.570 4	0.546 8	0.507 5	**0.595 5**
D14	0.726 3	0.773 8	0.696 6	0.773 8	0.558 9	0.570 6	**0.810 8**
D17	0.660 0	0.658 9	0.548 2	0.658 9	0.607 5	0.600 0	**0.698 2**
D18	0.630 0	0.680 0	0.600 0	0.680 0	0.540 0	—	**0.730 0**
D19	0.500 0	0.540 0	0.540 0	0.540 0	0.490 0	—	**0.560 0**
D21	0.933 3	0.883 3	0.950 0	0.900 0	**0.966 7**	—	**0.966 7**
D24	0.759 1	0.777 5	0.683 1	0.777 5	0.592 3	0.647 1	**0.803 6**
D26	0.889 9	0.913 0	0.916 4	0.913 0	0.809 9	0.809 9	**0.953 1**
D27	0.960 0	0.960 0	0.970 0	0.960 0	0.836 7	0.876 7	**1.000 0**
D29	0.952 5	0.943 9	0.949 4	0.943 9	0.881 8	0.896 9	**0.977 5**
D30	0.960 3	0.910 3	0.960 3	0.910 3	0.847 1	0.844 1	**0.969 1**

As can be seen from Table 4, AMF achieves the highest AUC values on 17 out of 30 datasets, overall outperforming the remaining six filter methods. AMF achieves significantly higher AUC values on D8 and D11 datasets compared with other algorithms. For D1, D7, and D24 datasets, the ReliefF algorithm performs worse than other filter methods, while AMF avoids the effects of the worse filter methods and effectively selects the better features, achieving the highest scores among all algorithms.

3.6 Performance analysis of adaptive feature selection

AMF-SGSK was compared with other eight algorithms, namely sampling-based methods: SMOTE and MWMOTE; metaheuristic-based methods: binary grey wolf optimization algorithm (BGWO)[51] and binary grasshopper optimization algorithm (BGOA)[52]; hybrid feature selection methods: ReliefF-GWO[53], RFACO-GS[54], IG-PSO[55], and IG-SSO[56]. Table 5 compares the *G*-mean of algorithms across all datasets, with bold indicating the highest scores among the nine methods.

Table 5 shows that AMF-SGSK achieves the highest *G*-mean values on 28 out of 30 datasets. AMF-SGSK attains the *G*-mean of 1 in D5, D7, and D29 datasets, while all other algorithms fail to match this performance. AMF-SGSK outperforms other methods on individual datasets, with *G*-mean that is 7.55% higher than the second-highest IG-SSO in D3 and 6.21% higher than the second-highest IG-PSO in D14. AMF-SGSK ranks second or third among the nine algorithms in datasets where it does not achieve the best performance. Fig. 4 summarizes the mean values of evaluation metrics across 30 datasets for each algorithm.

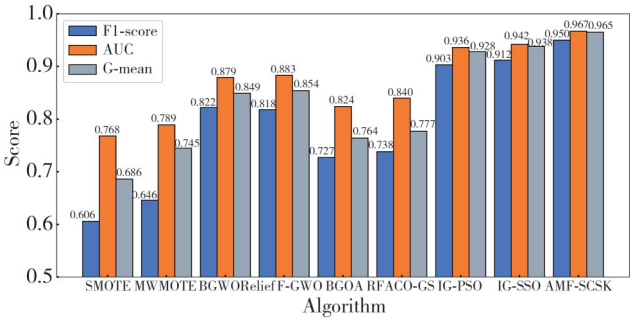

Fig. 4 Comparison of mean values across algorithms for different evaluation metrics

Fig. 4 demonstrates that the average values of *F*1-score, AUC, and *G*-mean of AMF-SGSK are 0.950, 0.967, 0.965, respectively. They are the highest among all algorithms. Its *G*-mean value outperforms IG-PSO, ReliefF-GWO, and BGOA by 3.7%, 11.1%, and 20.1%, respectively. AMF-SGSK overall shows superior classification performance, confirming the method demonstrates superiority in handling high-dimensional imbalanced data. Fig. 5 presents the average

rankings of the multiple evaluation metrics of each algorithm among 30 datasets.

Table 5 Performance comparison of algorithms for KNN classifier

ID	SMOTE	MWMOTE	BGWO	BGOA	ReliefF-GWO	RFACO-GS	IG-PSO	IG-SSO	AMF-SGSK
D1	0.824 8	0.823 0	0.921 9	0.861 0	0.746 2	0.745 1	0.880 2	0.881 3	**0.926 1**
D2	0.652 3	0.776 7	0.729 6	0.652 0	0.612 9	0.619 2	0.780 0	0.767 0	**0.787 3**
D3	0.437 8	0.424 1	0.731 8	0.532 4	0.663 7	0.663 7	0.789 4	0.799 4	**0.874 9**
D4	0.650 8	0.723 7	0.895 5	0.843 4	0.967 3	0.923 8	0.978 9	0.946 2	**0.987 1**
D5	0.975 5	0.981 0	0.997 3	0.970 8	0.997 3	0.960 6	0.970 8	0.997 3	**1.000 0**
D6	0.963 2	0.950 8	0.973 2	0.973 2	0.946 4	0.973 2	**1.000 0**	**1.000 0**	**1.000 0**
D7	0.965 7	0.984 4	0.624 3	0.707 1	0.541 4	0.482 8	0.882 8	0.882 8	**1.000 0**
D8	0.962 0	0.994 8	0.682 8	0.141 4	0.565 7	0.282 8	**1.000 0**	**1.000 0**	**1.000 0**
D9	0.860 2	0.826 2	0.950 4	0.898 8	0.926 6	0.963 3	0.935 5	0.972 2	**1.000 0**
D10	0.780 7	0.825 8	0.926 6	0.863 9	0.914 5	0.878 8	0.963 3	0.985 2	**1.000 0**
D11	0.633 3	0.658 6	0.896 1	0.624 4	0.901 9	0.721 9	0.952 9	0.985 2	**0.985 2**
D12	0.847 5	0.846 1	0.947 4	0.894 9	0.987 1	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**
D13	0.389 3	0.398 2	0.625 6	0.480 9	0.744 0	0.680 4	0.644 9	0.685 0	**0.789 8**
D14	0.509 7	0.470 9	0.702 6	0.615 9	0.739 5	0.738 4	0.893 4	0.890 0	**0.955 5**
D15	0.432 7	0.446 3	0.699 7	0.595 2	0.692 3	0.688 2	0.794 8	0.798 3	**0.840 4**
D16	0.795 7	0.780 7	0.936 7	0.912 7	0.978 9	0.936 7	0.988 6	**1.000 0**	**1.000 0**
D17	0.846 6	0.637 5	0.852 0	0.808 4	**0.935 2**	0.922 3	0.919 6	0.874 9	0.919 6
D18	0.000 0	0.671 4	0.751 2	0.341 4	0.824 3	0.682 8	0.979 5	0.979 5	**1.000 0**
D19	0.044 7	0.367 9	0.141 4	0.134 2	0.473 5	0.000 0	0.682 8	0.867 9	**0.941 4**
D20	0.000 0	0.081 6	0.978 7	0.945 9	0.799 4	0.624 6	0.965 1	0.982 6	**1.000 0**
D21	0.128 4	0.745 8	0.945 9	0.766 1	0.965 1	0.244 9	0.982 6	0.982 6	**1.000 0**
D22	0.984 3	0.985 2	**1.000 0**	0.987 1	0.999 1	0.987 1	**1.000 0**	**1.000 0**	**1.000 0**
D23	0.983 0	0.977 3	0.992 7	0.988 9	**0.992 7**	**0.992 7**	**0.992 7**	**0.992 7**	**0.992 7**
D24	0.581 2	0.574 0	0.813 8	0.740 5	0.917 8	0.905 6	0.931 7	0.901 8	**0.972 2**
D25	0.810 5	0.786 5	0.920 3	0.895 7	0.942 2	0.942 2	**1.000 0**	**1.000 0**	**1.000 0**
D26	0.822 4	0.861 3	0.936 7	0.888 5	0.909 7	0.888 7	0.978 9	0.978 9	**1.000 0**
D27	0.836 6	0.880 9	0.989 7	0.958 3	0.978 9	0.978 9	**1.000 0**	**1.000 0**	**1.000 0**
D28	0.969 7	0.969 7	0.982 6	0.982 6	0.982 6	0.982 6	0.982 6	**1.000 0**	**1.000 0**
D29	0.917 6	0.918 3	0.945 8	0.951 4	0.988 5	0.970 9	0.989 0	0.994 4	**1.000 0**
D30	0.972 3	0.981 9	0.970 2	0.970 2	0.973 2	0.914 6	0.967 2	**0.994 1**	0.973 2

Fig. 5 demonstrates that AMF-SGSK has the highest rankings for three average metrics, all of which are 1.167, significantly outperforming other algorithms. The next are IG-PSO and IG-SSO. AMF-SGSK outperforms two algorithms. The superiority may stem from two factors. On the one hand, AMF combines multiple superior filter methods, which indeed achieves better results than a single filter method. On the other hand, searching across multiple subspaces enables more effective identification of the optimal feature subset.

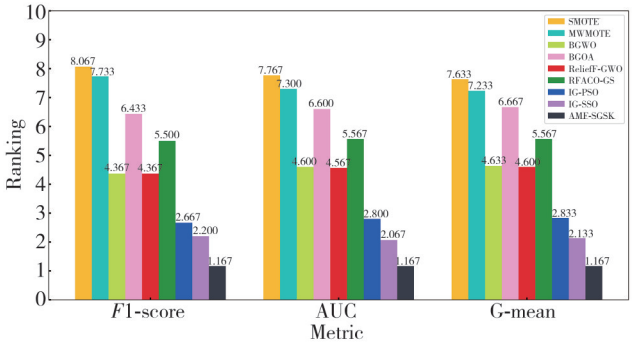

Fig. 5 Average ranking of evaluation metrics for each algorithm

IG-PSO and IG-SSO are ranked higher than ReliefF-GWO and RFACO-GS, which may be attributed to the fact that the filter method IG selected more critical features in the datasets used in this study, thereby achieving better performance during the optimization process. And the four hybrid feature selection methods generally outperform the meta-heuristic-only feature selection methods BGWO and BGOA, indicating that it is difficult to obtain the optimal feature subset for the single optimization algorithm in dealing with high-dimensional imbalanced data. SMOTE and MWMOTE achieve the lowest rankings among all methods. This indicates that traditional resampling approaches may not be effective for handling high-dimensional imbalanced data, as irrelevant and redundant features still exist and fail to improve classification performance. Table 6 represents the average selected features ratio to the total number of features for each algorithm in the selected dataset.

Table 6 Comparison of average feature selection ratio between AMF-SGSK and other five methods

ID	BGOA	ReliefF-GWO	RFACO-GS	IG-PSO	IG-SSO	AMF-SGSK
D13	0.321 6	0.049 6	0.039 5	0.026 0	0.032 9	**0.008 9**
D14	0.327 2	0.033 9	0.044 6	0.040 6	0.036 5	**0.004 1**
D16	0.522 9	0.035 2	0.048 1	0.048 4	0.038 0	**0.003 1**
D20	0.515 5	0.013 3	0.043 8	0.049 0	0.042 0	**0.006 2**
D21	0.445 6	0.005 3	0.028 4	0.052 7	0.052 7	**0.000 9**
D24	0.480 1	0.033 3	0.044 0	0.034 6	0.033 6	**0.001 1**
D25	0.467 4	0.041 8	0.044 8	0.051 3	0.046 7	**0.001 2**
D26	0.441 3	0.028 6	0.042 6	0.043 6	0.044 4	**0.007 7**
D28	0.476 1	0.039 4	0.044 5	0.057 5	0.041 0	**0.004 8**
D29	0.327 4	0.031 2	0.031 2	0.017 6	0.027 1	**0.001 7**

Table 6 shows that AMF-SGSK achieves feature selection ratio with less than 0.01, which significantly

lower than that of other methods. Among the other comparison methods, the ratios of four hybrid feature selection methods range mostly between 0.03—0.05, while BGOA shows the highest ratios. This indicates that AMF-SGSK may remove more irrelevant and redundant features during the filter stage compared with other hybrid methods, thereby reducing the dimensionality to a lower level. This also facilitates more effective dimensionality reduction in the optimization stage. For D21, D22, and D24 datasets with over 10 000 dimension, AMF-SGSK achieves extremely low ratios that are 0.000 9, 0.001 1, and 0.001 2, respectively, which means that only about 10 features are retained at the end. This indicates that AMF-SGSK adaptively selects fewer features and still maintains a high classification accuracy in the data with a lot of irrelevant and redundant features.

3.7 Visualization of experimental results

To visualize the effect of AMF-SGSK, the features selected by the algorithm are reduced to two dimensions using t-SNE, enabling clear visualization of sample distributions. The two-dimensional projection is visualized in Fig.6 through a scatter plot, with red and blue points representing majority and minority class samples, respectively.

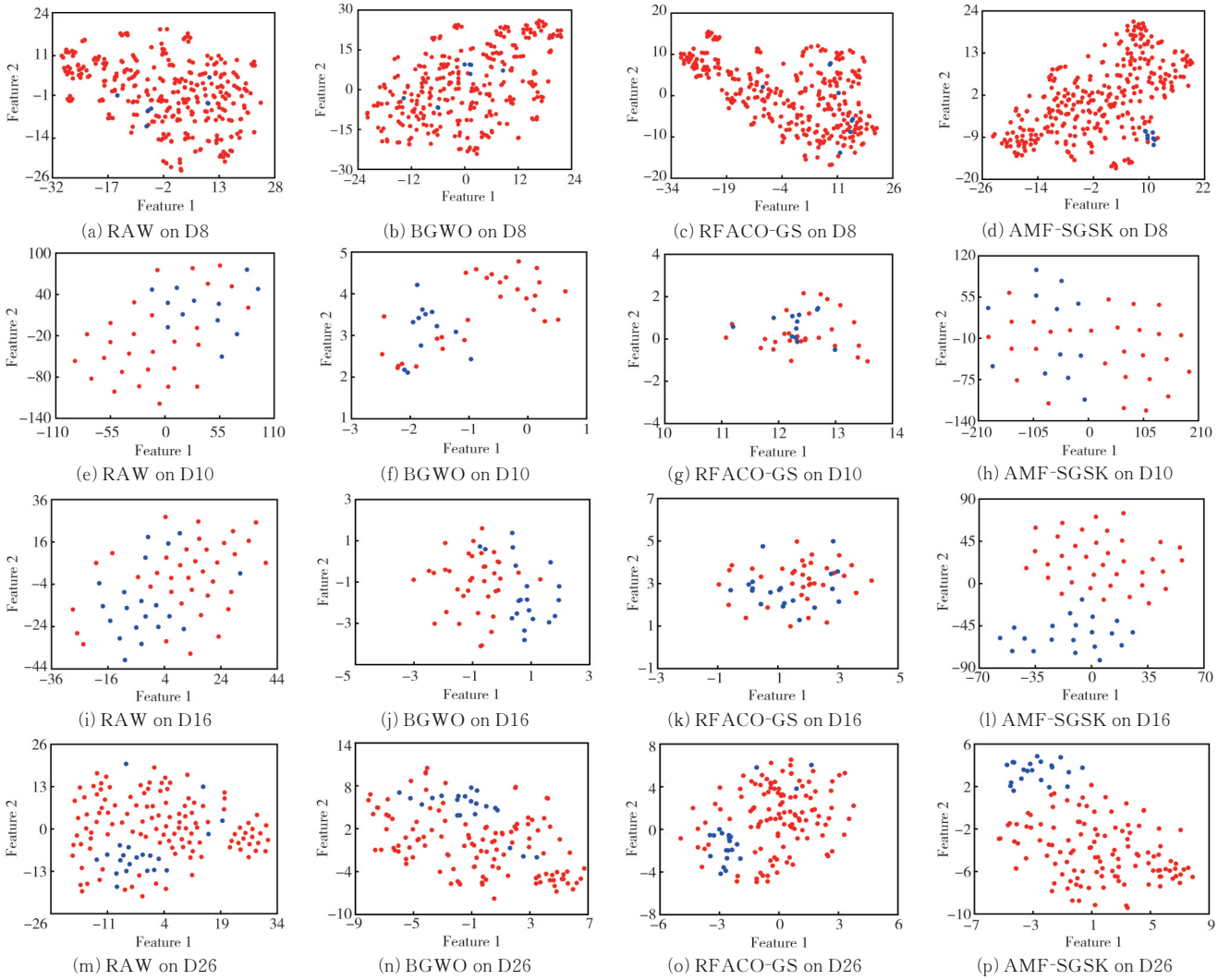

Fig. 6 Visualization comparison between ours and other methods

Fig.6 reveals that the D8 dataset exhibits high imbalance ratio, where the minority class samples in the original data are dispersed and overlap with majority class samples, almost regarded as outliers, which makes it difficult to identify the minority class from the majority class. After selecting the more important features through AMF-SGSK, the blue points form a distinct cluster near the majority class boundary, significantly improving separability. In contrast, the minority class samples of BGWO and RFACO-GS remain scattered. For the D10 dataset, AMF-SGSK maintains clear decision boundaries compared to the original data, while BGWO and RFACO-

GS produce more ambiguous distributions that significantly hinder minority class identification. For the D16 and D26 datasets, while the original data show minority class samples embedded within the majority class region with scattered outliers, the minority class samples are clustered together without outliers and produce a clear decision boundary through AMF-SGSK, enabling the samples to be easily identified.

4 Conclusions

In this paper, we proposed a novel adaptive feature selection method named AMF-SGSK for high-dimensional imbalanced data classification. The method used adaptive multi-filter AMF to adaptively remove a large number of irrelevant and redundant features in high-dimensional imbalanced data based on the proposed feature hardness, and then used subspace-based gaining sharing knowledge-based algorithm SGSK to more effectively obtain the feature subset with the best classification performance, which improved the classification accuracy of high-dimensional imbalanced data. Comparing AMF-SGSK with eight other algorithms on 30 benchmark high-dimensional imbalanced data, the results show that AMF-SGSK achieved the highest G-mean values in 28 datasets. Additionally, it obtained the top average rankings in $F1$-score, AUC, and G-mean metrics. The experimental results demonstrate that AMF-SGSK significantly outperformed the competing algorithms and effectively improved classification accuracy for high-dimensional imbalanced data. In this paper, we studied the binary classification problem of high-dimensional imbalanced data. In multi-class problems, the more complex feature space structures and class interaction relationships make effective feature selection important for enhancing classification accuracy of multi-class data and reducing model complexity. Therefore, future research will focus on investigating the multi-class classification of high-dimensional imbalanced data.

Acknowledgement

This work was supported by Fundamental Research Program of Shanxi Province (Nos. 202203021211088, 202403021212254, 202403021221109), and Graduate Research Innovation Project in Shanxi Province (No. 2024KY616).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] QIU Y, MA L, PRIYADARSHI R. Deep learning challenges and prospects in wireless sensor network deployment. Archives of Computational Methods in Engineering, 2024, 31(6): 3231-3254.

[2] LUO T, XIE J P, ZHANG B T, et al. An improved levy chaotic particle swarm optimization algorithm for energy-efficient cluster routing scheme in industrial wireless sensor networks. Expert Systems with Applications, 2024, 241: 122780.

[3] PRIYADARSHI R. Exploring machine learning solutions for overcoming challenges in IoT-based wireless sensor network routing: a comprehensive review. Wireless Networks, 2024, 30(4): 2647-2673.

[4] OUADERHMAN T, CHAMLAL H, JANANE F Z. A new filter-based gene selection approach in the DNA microarray domain. Expert Systems with Applications, 2024, 240: 122504.

[5] KAMALOV F, LEUNG H H. Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 2020, 19(1): 2040013.

[6] CUI J Y, ZONG L S, XIE J H, et al. A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data. Applied Intelligence, 2023, 53(1): 272-288.

[7] HE H B, GARCIA E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[8] THAI-NGHE N, GANTNER Z, SCHMIDT-THIEME L. Cost-sensitive learning methods for imbalanced data// The 2010 International Joint Conference on Neural Networks, July 18-23, 2010, Barcelona, Spain. New York: IEEE, 2010: 1-8.

[9] SAGI O, ROKACH L. Ensemble learning: a survey. WIREs Data Mining and Knowledge Discovery, 2018, 8(4): e1249.

[10] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[11] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning//Advances in Intelligent Computing. Berlin, Heidelberg: Springer, 2005: 878-887.

[12] HE H B, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning//2008 IEEE International Joint Conference on Neural Networks, June 1-8, 2008, Hong Kong, China. New York: IEEE, 2008: 1322-1328.

[13] RAMENTOL E, CABALLERO Y, BELLO R, et al. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and Information Systems, 2012, 33(2): 245-265.

[14] BARUA S, ISLAM M M, YAO X, et al. MWMOTE:

majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405-425.

[15] KHALID S, KHALIL T, NASREEN S. A survey of feature selection and feature extraction techniques in machine learning//2014 Science and Information Conference, August 27-29, 2014, London, UK. New York: IEEE, 2014: 372-378.

[16] SUN L, SI S S, DING W P, et al. TFSFB: Two-stage feature selection via fusing fuzzy multi-neighborhood rough set with binary whale optimization for imbalanced data. Information Fusion, 2023, 95: 91-108.

[17] SONG X F, ZHANG Y, GONG D W, et al. A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data. IEEE Transactions on Cybernetics, 2022, 52(9): 9573- 9586.

[18] LIU H, SETIONO R. Chi2: feature selection and discretization of numeric attributes//7th IEEE International Conference on Tools with Artificial Intelligence, November 5-8, 1995, Herndon, VA, USA. New York: IEEE, 1995: 388-391.

[19] SHANG C X, LI M, FENG S Z, et al. Feature selection via maximizing global information gain for text classification. Knowledge-Based Systems, 2013, 54: 298-309.

[20] PENG H C, LONG F H, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[21] ROBNIK-ŠIKOJA M, KONONENKO I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning, 2003, 53(1): 23-69.

[22] DOKEROGLU T, DENIZ A, KIZILOZ H E. A comprehensive survey on recent metaheuristics for feature selection. Neurocomputing, 2022, 494: 269-296.

[23] WHITLEY D. A genetic algorithm tutorial. Statistics and Computing, 1994, 4(2): 65-85.

[24] EBERHART, SHI Y H. Particle swarm optimization: developments, applications and resources//2001 Congress on Evolutionary Computation, May 27-30, 2001, Seoul, Korea. New York: IEEE, 2001: 81-86.

[25] DORIGO M, BIRATTARI M, STUTZLE T. Ant colony optimization. IEEE Computational Intelligence Magazine, 2006, 1(4): 28-39.

[26] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer. Advances in Engineering Software, 2014, 69: 46-61.

[27] SAREMI S, MIRJALILI S, LEWIS A. Grasshopper optimisation algorithm: theory and application. Advances in Engineering Software, 2017, 105: 30-47.

[28] YIN Y H, JANG-JACCARD J, XU W, et al. IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. Journal of Big Data, 2023, 10(1): 15.

[29] ZHAO B J, YANG D S, KARIMI H R, et al. Filter-wrapper combined feature selection and adaboost-weighted broad learning system for transformer fault diagnosis under imbalanced samples. Neurocomputing, 2023, 560: 126803.

[30] VOMMI A M, BATTULA T K. A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: a COVID-19 case study. Expert Systems with Applications, 2023, 218: 119612.

[31] XU Y H, YU Z W, PHILIP CHEN C L P. Classifier ensemble based on multiview optimization for high-dimensional imbalanced data classification. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 870-883.

[32] ZHANG C K, ZHOU Y, GUO J W, et al. Research on classification method of high-dimensional class-imbalanced datasets based on SVM. International Journal of Machine Learning and Cybernetics, 2019, 10(7): 1765-1778.

[33] AYDOGAN E K, OZMEN M, DELICE Y. CBR-PSO: cost-based rough particle swarm optimization approach for high-dimensional imbalanced problems. Neural Computing and Applications, 2019(10), 31: 6345-6363.

[34] ZHANG Y, WANG Y H, GONG D W, et al. Clustering-guided particle swarm feature selection algorithm for high-dimensional imbalanced data with missing values. IEEE Transactions on Evolutionary Computation, 2021, 26(4): 616-630.

[35] ALMOTAIRI K H. Gene selection for high-dimensional imbalanced biomedical data based on marine predators algorithm and evolutionary population dynamics. Arabian Journal for Science and Engineering, 2024, 49(3): 3935-3961.

[36] MOAYEDIKIA A, ONG K L, BOO Y L, et al. Feature selection for high dimensional imbalanced class data using harmony search. Engineering Applications of Artificial Intelligence, 2017, 57: 38-49.

[37] ABDULRAUF SHARIFAI G, ZAINOL Z. Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm. Genes, 2020, 11(7): 717.

[38] SHARIFAI A G, ZAINOL Z B. Multiple filter-based rankers to guide hybrid grasshopper optimization algorithm and simulated annealing for feature selection with high dimensional multi-class imbalanced datasets. IEEE Access, 2021, 9: 74127-74142.

[39] SAHU B, PANIGRAHI A, ROUT S K, et al. Hybrid multiple filter embedded political optimizer for feature selection//2022 International Conference on Intelligent Controller and Computing for Smart Power, July 21-23, 2022, Hyderabad, India. New York: IEEE, 2022: 1-6.

[40] LIU Y, WANG Y Z, REN X G, et al. A classification method based on feature selection for imbalanced data. IEEE Access, 2019, 7: 81794-81807.

[41] KIM J, KANG J, SOHN M. Ensemble learning-based filter-centric hybrid feature selection framework for high-

dimensional imbalanced data. Knowledge-Based Systems, 2021, 220: 106901.
[42] ST L, WOLD S. Analysis of variance (ANOVA). Chemometrics and Intelligent Laboratory Systems, 1989, 6(4): 259-272.
[43] SAIDI R, BOUAGUEL W, ESSOUSSI N. Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. Machine Learning Paradigms: Theory and Application. Cham: Springer International Publishing, 2019: 3-24.
[44] SENLIOL B, GULGEZEN G, YU L, et al. Fast Correlation Based Filter (FCBF) with a different search strategy//2008 23rd International Symposium on Computer and Information Sciences, October 27-29, 2008, Istanbul, Turkey. New York: IEEE, 2008: 1-4.
[45] MOHAMED A W, HADI A A, MOHAMED A K. Gaining-sharing knowledge based algorithm for solving optimization problems: a novel nature-inspired algorithm. International Journal of Machine Learning and Cybernetics, 2020, 11(7): 1501-1529.
[46] COVER T, HART P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.
[47] VANSCHOREN J, VAN RIJN J N, BISCHL B, et al. OpenML: networked science in machine learning. ACM SIGKDD Explorations Newsletter, 2014, 15(2): 49-60.
[48] FRANK A. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2010.
[49] LI J D, CHENG K W, WANG S H, et al. Feature selection: a data perspective. ACM Computing Surveys, 2017, 50(6): 1-45.
[50] WARDHANI N W S, ROCHAYANI M Y, IRIANY A, et al. Cross-validation metrics for evaluating classification performance on imbalanced data//2019 International Conference on Computer, Control, Informatics and its Applications, October 23-24, 2019, Tangerang, Indonesia. New York: IEEE, 2019: 14-18.
[51] EMARY E, ZAWBAA H M, HASSANIEN A E. Binary grey wolf optimization approaches for feature selection. Neurocomputing, 2016, 172: 371-381.
[52] MAFARJA M, ALJARAH I, FARIS H, et al. Binary grasshopper optimisation algorithm approaches for feature selection problems. Expert Systems with Applications, 2019, 117: 267-286.
[53] DASS S, MISTRY S, SARKAR P. Identification of promising biomarkers in cancer diagnosis using a hybrid model combining reliefF and grey Wolf optimization//International Conference on Communication and Intelligent Systems. Singapore: Springer Nature Singapore, 2022: 311-321.
[54] SUN L, KONG X L, XU J C, et al. A hybrid gene selection method based on ReliefF and ant colony optimization algorithm for tumor classification. Scientific Reports, 2019, 9(1): 8978.
[55] YİĞİT F, BAYKAN Ö K. A new feature selection method for text categorization based on information gain and particle swarm optimization//2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, November 27-29, 2014, Shenzhen, China. New York: IEEE, 2014: 523-529.
[56] YE C C, PAN J L, JIN Q. An improved SSO algorithm for cyber-enabled tumor risk analysis based on gene selection. Future Generation Computer Systems, 2019, 92: 407-418.

高维不平衡数据分类的自适应特征选择方法

吴建臻[1], 薛　震[1,2*], 张亮亮[1], 杨　旭[1]

1. 中北大学 数学学院, 山西 太原 030051;
2. 香港城市大学 数学系, 香港 九龙 999077

摘　要： 网络安全和生物医学领域的数据经常呈现高维度与类不平衡的特点。为解决高维不平衡数据中不相关和冗余特征多，导致少数类分类精度低的问题，本文提出一种基于自适应多过滤法与子空间知识获取共享的特征选择方法（AMF-SGSK）。首先，对数据进行欠采样以获得平衡数据。其次，将各过滤法的特征得分与AUC得分相结合，提出了一种评判特征重要性的概念，特征硬度，据此可自适应地选择出重要特征。最后，通过在多个子空间使用知识获取共享算法，得到最佳特征子集，以对高维不平衡数据进行有效降维。将AMF-SGSK与BGWO、IG-SSO等8种经典算法，在30个公开数据集上依据$F1$-score、AUC与G-mean指标进行比较，发现AMF-SGSK算法性能良好。AMF-SGSK的$F1$-score、AUC与G-mean的平均值分别是0.950, 0.967, 0.965，均为所有算法中的最高值，其G-mean平均值比IG-PSO、ReliefF-GWO、BGOA的分别高出3.72%, 11.12%和20.06%。并且在所选的10个数据集上，特征选择比率均小于0.01，优于其他算法。在可视化分析中，AMF-SGSK算法优于其他算法，能够使少数类样本更容易被识别出。AMF-SGSK能自适应地去除不相关和冗余特征，有效提升高维不平衡数据的分类精度，可为相关领域提供科学技术参考。

关键词： 高维不平衡数据；自适应特征选择；自适应多过滤法；特征硬度；知识获取共享算法；元启发式算法

引用格式： WU Jianzhen, XUE Zhen, ZHANG Liangliang, et al. Adaptive feature selection method for high-dimensional imbalanced data classification. Journal of Measurement Science and Instrumentation, 2025, 16(4): 612-624. DOI: 10.62756/jmsi.1674-8042.2025059