Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 6, No. 3, 2026

A Feature Selection Method for High-Dimensional Medical Data Based on Adaptive Redundancy Penalty

Zengguang Wang *

Henan Polytechnic University, Jiaozuo 454000, China

Abstract: High-dimensional medical data are often characterized by extremely high feature dimensionality, limited sample sizes, substantial redundant information, and strong noise interference. To address these challenges, this paper proposes a filter-based feature selection method, termed ARMR (Adaptive Redundancy-based Minimum Redundancy Maximum Relevance Feature Selection). The proposed method first employs mutual information to select initial features from the original feature set. It then uses joint mutual information to characterize the joint discriminative capability of candidate features and selected features with respect to class labels. Furthermore, an adaptive penalty factor, constructed from conditional mutual information and three-way interaction information, is introduced to dynamically weight the redundancy term, thereby enabling a more flexible balance between feature complementarity and redundancy suppression. To evaluate the effectiveness of the proposed method, experiments were conducted on six benchmark datasets and four medical datasets, with comparisons against several representative and recently developed methods, including CFR, DCSF, JMI, MRMD, and mRMR. Two classifiers, namely SVM and NB, were employed for performance assessment. The results show that the proposed method achieves superior overall performance. In particular, it attains the highest average classification accuracy of 84.58% with SVM and 76.36% with NB. Further analysis based on classification accuracy curves under different feature subset sizes, together with boxplots of F1-score and AUC, demonstrates that the proposed method exhibits strong performance in terms of classification accuracy, stability, and classifier adaptability.

Keywords: High-dimensional Medical Data, Feature Selection, Joint Mutual Information, Adaptive Redundancy Penalty, Classification Accuracy.

1. Introduction

With the rapid development of information technology and sequencing techniques in medical data processing, a large amount of high-dimensional data has been generated in fields such as biomedicine, clinical diagnosis, and intelligent decision support [1]. These data are typically characterized by extremely high feature dimensionality, limited sample sizes, substantial redundancy, and strong noise interference [2]. Directly feeding all original features into classification models often leads to the curse of dimensionality, overfitting, reduced training efficiency, and unstable classification performance. Therefore, preprocessing of high-dimensional data has become a critical issue in intelligent diagnosis and medical data analysis [3].

Feature selection is one of the key techniques in data preprocessing and pattern recognition [4]. Its main objective is to select a subset of features that is highly relevant to the target task while containing as little redundancy as possible, so as to reduce dimensionality and preserve discriminative information [5]. Effective feature selection can improve the accuracy, stability, and generalization ability of classification models, reduce computational cost during training and prediction, and enhance the interpretability of results, thereby providing support for biomarker discovery and clinical decision-making.

According to the search strategy and the way they interact with classifiers, feature selection methods can generally be divided into wrapper, embedded, and filter methods. Wrapper methods evaluate candidate feature subsets repeatedly using a classifier and select the subset with the best classification performance. Although they usually achieve good performance, their computational cost is high, which limits their applicability to high-dimensional data. Representative wrapper methods include recursive feature elimination [6] and simulated annealing algorithms [7]. Embedded methods incorporate feature selection into the model training process, achieving a certain trade-off between efficiency and performance; however, they usually depend on specific classifier structures and thus have limited generality. Typical embedded methods include L1-norm-based methods [8] and ridge regression [9]. In contrast, filter methods evaluate and rank features according to the statistical properties of the data itself without relying on a specific classifier. Due to their simplicity, high efficiency, and suitability for high-dimensional small-sample data, filter methods have been widely studied in medical data analysis [10]. Common filter methods include Pearson correlation coefficient-based methods [11] and mutual information-based methods [12].

Among filter-based methods, information-theoretic approaches have attracted considerable attention because they characterize feature effectiveness from the perspectives of uncertainty reduction and statistical dependency [5]. Information measures such as mutual information, conditional mutual information, and three-way interaction information provide effective tools for quantifying both feature relevance to class labels and redundancy among features [13], [14]. Based on criteria such as maximal relevance, minimal redundancy, conditional dependency, and higher-order information interaction, many information-theoretic filter methods have been proposed and successfully applied to high-dimensional classification tasks. For instance, Peng et al. [15] proposed the minimal-redundancy-maximal-relevance (mRMR) algorithm, which balances relevance and redundancy during feature selection. Yang et al. [16] introduced the Joint Mutual Information (JMI) algorithm,

which evaluates the class-related contribution of candidate features using joint mutual information. Bennasar et al. [17] further proposed the Joint Mutual Information Maximization (JMIM) algorithm, which adopts a max–min strategy for feature selection.

Despite their effectiveness, existing information-theoretic filter methods still have several limitations. First, although most methods consider both feature relevance and feature redundancy, redundancy suppression is usually implemented through a fixed penalty form, which lacks flexibility in adapting to different information relationships between candidate and selected features. Second, while some methods incorporate joint mutual information or conditional mutual information to improve evaluation accuracy, they still lack an effective mechanism to dynamically balance the complementary discriminative information contributed by candidate and selected features against their redundancy. Third, as the selected feature subset grows, fixed redundancy penalties may either over-suppress informative complementary features or under-penalize highly redundant ones, thereby affecting the compactness of the selected subset and the final classification performance.

To address these issues, this paper proposes an adaptive filter-based feature selection algorithm based on joint mutual information, namely ARMR (Adaptive Redundancy-based Minimum Redundancy Maximum Relevance Feature Selection). The main contributions of this work are summarized as follows:

(1) Joint mutual information is introduced to evaluate the contribution of candidate features together with selected features to the class label, replacing the traditional mutual information-based relevance evaluation.

(2) An adaptive redundancy penalty mechanism is developed to dynamically adjust the weight of the redundancy term according to the information relationships among candidate features, selected features, and class labels, thereby providing a more flexible estimation of feature redundancy.

(3) A new filter-based feature evaluation criterion is established under the maximum relevance and minimum redundancy framework. The proposed method retains the efficiency of traditional filter methods while achieving a more flexible balance between joint discriminative ability and redundancy suppression, which is expected to improve feature selection quality and subsequent classification performance on high-dimensional medical data.

The remainder of this paper is organized as follows. Section 2 introduces the commonly used information-theoretic measures. Section 3 presents the proposed method. Section 4 reports the comparative experiments and corresponding analysis. Finally, Section 5 concludes the paper and outlines future work.

2. Introduction to Related Theoretical Knowledge

In this section, we briefly review several commonly used information-theoretic measures. Information entropy is used to quantify the uncertainty of a random variable. Let X be a discrete random variable with sample space $\mathcal{X}$ and probability distribution $p(x)$. Then, the information entropy of X is defined as follows:

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \quad (1)$$

For two discrete random variables X and Y, the joint entropy is defined as:

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \quad (2)$$

Conditional entropy represents the remaining uncertainty of one random variable when another random variable is given. The conditional entropy of X given $Y$is defined as:

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \quad (3)$$

Mutual information is used to measure the amount of shared information between two random variables, that is, the extent to which one variable can reduce the uncertainty of the other. The mutual information between X and Y is defined as:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \quad (4)$$

Conditional mutual information measures the remaining dependency between two random variables under the condition of a third random variable. For random variables X, Y and Z, the conditional mutual information is defined as:

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z) \quad (5)$$

Joint mutual information is used to measure the amount of information shared jointly by multiple random variables with another random variable. For random variables X, Y and Z, the joint mutual information is defined as:

$$I(X,Y;Z) = H(Z) - H(Z \mid X,Y) \quad (6)$$

For high-dimensional small-sample medical data, filter-based methods remain of greater practical value because they do not rely on classifier training, are easy to implement, and are computationally efficient. Therefore, how to further improve the rationality of candidate feature evaluation within the filter-based framework remains an important direction of current research.

3. The Proposed Method

$D = \{(x_1,c_1),(x_2,c_2),\dots,(x_N,c_N)\}$, stands for the original dataset, where N denotes the number of samples, $x_i = \{f_{i1},f_{i2},\dots,f_{iM}\}$ represents the feature vector of the i-th sample, M is the total number of original features, and c_i denotes the corresponding class label. The original feature set is denoted as:$F = \{f_1,f_2,\dots,f_M\}$.

The objective of feature selection is to select a feature subset of size K from the original feature set, $S = \{f_{s_1},f_{s_2},\dots,f_{s_K}\}, S \subseteq F$, Such that the selected subset can preserve as much discriminative information related to the class label as possible while reducing redundancy among features, thereby improving the classification performance of subsequent classifiers.

For ease of subsequent derivation, in this paper, X denotes a candidate feature, Y denotes a selected feature, Z denotes the class label, S denotes the current selected feature subset, and $|S|$ denotes the number of selected features.

3.1. ARMR Proposal Process

The joint discriminative ability of a candidate feature X and a selected feature Y with respect to the class label Z can be characterized by $I(Z;X,Y)$:

$$I(Z;X,Y) = I(Z;X) + I(Z;Y \mid X) \quad (7)$$

This quantity represents the amount of information shared jointly by the candidate feature, the selected feature, and the

class label. A larger value indicates that X and Y provide richer discriminative information for identifying the class label when considered together.

The information redundancy between a candidate feature and a selected feature can be measured by the mutual information, $I(X;Y)$. A larger value of $I(X;Y)$ indicates that the candidate feature and the selected feature share more overlapping information. If such a feature is still included in the selected subset, it may increase the internal redundancy of the feature subset and consequently affect its compactness and classification performance.

Therefore, when evaluating a candidate feature, it is necessary to impose a certain penalty on the mutual information between the candidate feature and the selected feature, so as to suppress redundant features from being selected into the feature subset.

To enable the redundancy penalty strength to vary dynamically according to the actual information relationship between features, this paper defines an adaptive penalty factor $\lambda(\mathrm{X},\mathrm{Y})$ as follows:

$$\lambda(\mathrm{X},\mathrm{Y}) = \begin{cases} \frac{I(\mathrm{X},\mathrm{Y}|Z)}{I(\mathrm{X};\mathrm{Y};Z)}, & I(\mathrm{X};\mathrm{Y};Z) > \varepsilon \\ 1, & I(\mathrm{X};\mathrm{Y};Z) \le \varepsilon \end{cases} \tag{8}$$

Where $\mathrm{i}(\mathrm{X};\mathrm{Y}|\mathrm{Z})$ denotes the conditional mutual information between the candidate feature and the selected feature given the class label, $\mathrm{I}(\mathrm{X};\mathrm{Y};\mathrm{Z})$ denotes the three-way interaction information, and ε is a very small positive constant introduced to avoid numerical instability when $\mathrm{I}(\mathrm{X};\mathrm{Y}|\mathrm{Z})$ is extremely small or close to zero.

The design rationale of this penalty factor is as follows.

(1) When $\mathrm{I}(\mathrm{X};\mathrm{Y};\mathrm{Z}) \ge \mathrm{I}(\mathrm{X};\mathrm{Y}|\mathrm{Z})$, it indicates that the candidate feature and the selected feature can form a relatively strong information interaction with the class label when considered jointly, suggesting that they may possess good complementarity for the classification task. In this case, $\lambda(\mathrm{X},\mathrm{Y})$ tends to be relatively small, which weakens the penalty imposed on the redundancy term $I(X;Y)$, thereby avoiding excessive suppression of candidate features with complementary discriminative ability.

(2) When $\mathrm{I}(\mathrm{X};\mathrm{Y};\mathrm{Z}) < \mathrm{I}(\mathrm{X};\mathrm{Y}|\mathrm{Z})$ and $\mathrm{I}(\mathrm{X};\mathrm{Y}|\mathrm{Z})$ remains relatively large, it indicates that the candidate feature and the selected feature still preserve a strong conditional dependency under the constraint of the class label, implying that they are more likely to contain overlapping information. In this case, $\lambda(\mathrm{X},\mathrm{Y})$ becomes relatively large, and a stronger penalty is imposed on the redundancy term, thereby enhancing the ability of the algorithm to suppress redundant features.

(3) When $\mathrm{I}(\mathrm{X};\mathrm{Y};\mathrm{Z})$ is very small or close to zero, the three-way interaction information cannot provide a stable basis for adjusting the redundancy penalty strength. To ensure the robustness of the evaluation function, $\lambda(\mathrm{X},\mathrm{Y}) = 1$ is adopted in this paper, under which the algorithm degenerates to the baseline redundancy penalty form.

Based on the above analysis, the comprehensive evaluation function of a candidate feature X with respect to the current selected feature subset S is defined as:

$$J(X) = \frac{1}{|S|}\sum_{Y\in S}[I(Z;X,Y) - \lambda(X,Y)I(X;Y)] \tag{9}$$

Where the first term, $I(Z;X,Y)$, represents the joint discriminative ability of the candidate feature and the selected feature with respect to the class label, and the second term, $\lambda(X,Y)I(X;Y)$ denotes the adaptively weighted redundancy penalty term.

By calculating this quantity for each feature in the current selected subset and then taking the average, the scale bias caused by the increasing number of selected features during the iterative process can be alleviated to some extent, making the evaluation values at different stages of iteration more comparable. A larger value of $J(X)$ indicates that, under the current selected feature subset, the candidate feature can provide stronger joint discriminative information while exhibiting less redundancy. Therefore, such a feature is more suitable for inclusion in the selected feature subset.

3.2. ARMR Feature Selection Process

In the initial stage, since the selected feature subset is empty, the evaluation function (9) cannot be directly used to assess candidate features. Therefore, this paper first adopts the maximum relevance principle and selects, from the original feature set, the feature with the largest mutual information with the class label as the initial feature, i.e.,

$$f^{(1)} = arg\ \max_{f_i\in F} I\ (f_i;Z) \tag{10}$$

This strategy ensures that the first feature included in the selected subset has strong discriminative capability with respect to the class label, thereby providing a solid foundation for the subsequent iterative selection process.

Algorithm 1 ARMR

Input: dataset D, original feature set F, class label set C, target number of selected features K

Output: feature subset S.

1. Initialize the feature subset S as an empty set.
2. Compute the mutual information between each feature and the class label C.
3. Select the feature with the largest mutual information with the class label as the initial feature, and add it to S.
4. Precompute the information measures between any pair of features and their corresponding information quantities related to the class label.
5. While $|S| < K$, repeat the following steps:
6. For each unselected candidate feature f_i, calculate its comprehensive evaluation score $J(f_i)$.
7. Rank all candidate features in descending order according to their evaluation scores, and select the best feature f^*.
8. Add the feature f^* to the current feature subset S.
9. If $|S|$ has not yet reached K, continue the iteration.
10. End the loop.
11. Output the final feature subset S.

Let the current selected feature subset be $S = \{s_1, s_2, \dots, s_t\}$, where t denotes the number of features that have been selected so far. For any unselected candidate feature $f_i \in F\backslash S$, its comprehensive evaluation score is calculated as:

$$J(f_i) = \frac{1}{|S|}\sum_{s_j\in S}\left[I(Z;f_i,s_j) - \lambda(f_i,s_j)I(f_i;s_j)\right] \tag{11}$$

Where

$$\lambda(f_i,s_j) = \begin{cases} \frac{I(f_i;s_j|Z)}{I(f_i;s_j;Z)}, & I(f_i;s_j;Z) > \varepsilon \\ 1, & I(f_i;s_j;Z) \le \varepsilon \end{cases} \tag{12}$$

Then, the feature with the largest comprehensive evaluation score is selected from all candidate features and added to the selected feature subset, i.e.,

$$f^{(t+1)} = arg \max_{f_i \in F \setminus S} J(f_i) \quad (13)$$

The above procedure is repeated until the preset number of features K is reached or other stopping criteria are satisfied. In this study, for the sake of fair comparison among different methods, the feature selection process is terminated using a fixed number of selected features.

Based on the above feature evaluation criterion, this paper proposes a filter-based feature selection algorithm based on joint mutual information and adaptive redundancy penalty, ARMR (Adaptive Redundancy-based Minimum Redundancy Maximum Relevance Feature Selection). The detailed procedure is described in Algorithm 1.

4. Experimental Results and Analysis

4.1. Experimental Setup

To evaluate the effectiveness of the proposed method, six benchmark datasets and four medical datasets were selected for comparative experiments, and the results were compared with those of CFR [13], DCSF [14], JMI [18], MRMD [19], and mRMR [15]. Among them, ALLAML, ColonX, Prostate-GE, and TOX-171 are medical datasets, while the remaining datasets were obtained from the ASU feature selection repository [20] and the UCI Machine Learning Repository [21], [22]. All datasets are single-label classification datasets, covering medical diagnosis, image recognition, and other typical high-dimensional classification scenarios. They differ in terms of sample size, feature dimensionality, and number of classes. Detailed information on these datasets is provided in Table 1.

For continuous-valued features, an equal-frequency discretization method was adopted to transform them into discrete features, with each feature divided into five intervals. For datasets containing missing values, mean imputation was applied to continuous features, while mode imputation was used for discrete features, in order to ensure data integrity and reduce the influence of noise on the experimental results. In the experiments, all algorithms were evaluated on each dataset using ten-fold cross-validation, and the same training/test splits were maintained across different methods to ensure fairness. In addition, all methods were compared using a fixed number of selected features, with the number of selected features set to $K = 30$. Under each experimental setting, the procedure was repeated 10 times, and the average classification accuracy together with its standard deviation was reported.

Table 1. Description of datasets

Data sets	Instances	Features	Classes
Isolet	1560	617	26
Lsvt	126	310	2
Movement_libras	360	90	15
WarpAR10P	130	2400	10
WarpPIE10P	210	2420	10
Yale_32x32	165	1024	15
ALLAML	72	7129	2
ColonX	62	2000	2
Prostate-GE	102	5966	2
TOX-171	171	5748	4

Table 2 and Table3 present the experimental results of all methods under the support vector machine (SVM) and naive Bayes (NB) classifiers, respectively. All results are reported in the form of mean ± standard deviation, and the best results are highlighted in bold. The row labeled "Average" gives the average classification accuracy of each method across all datasets, while "W/T/L" denotes the numbers of wins/ties/losses of the proposed ARMR method compared with each competing method over different datasets.

Table 2 reports the comparison of average classification accuracy (mean ± std.) under the SVM classifier. It can be observed that the proposed method achieves an average accuracy of 84.58% over the ten datasets, which is the highest among all compared methods and clearly outperforms CFR (75.21%), DCSF (81.98%), JMI (81.25%), MRMD (82.32%), and mRMR (82.09%). In terms of overall average performance, the proposed method improves the classification accuracy by 9.37, 2.60, 3.33, 2.26, and 2.49 percentage points over CFR, DCSF, JMI, MRMD, and mRMR, respectively. These results demonstrate that the proposed method has stronger overall classification capability under the SVM classifier.

From the results on individual datasets, the proposed method achieves the best accuracy on seven datasets, namely Isolet, Lsvt, Movement_libras, WarpAR10P, WarpPIE10P, Yale, and TOX-171, showing a strong competitive advantage. In particular, on the Isolet and TOX-171 datasets, the proposed method attains accuracies of 85.80% and 85.66%, respectively, showing clear improvements over most competing methods. On multiclass or structurally complex datasets such as Movement_libras and WarpAR10P, the proposed method also obtains the best results, indicating that it can effectively identify more discriminative feature subsets from the original feature space for classification tasks. For face image datasets such as WarpPIE10P and Yale, the proposed method likewise achieves the highest or near-highest classification accuracy, further confirming its good generalization ability.

It can also be seen that the proposed method does not achieve the best performance on ALLAML, ColonX, and Prostate-GE, suggesting that its advantage is not absolute on some high-dimensional small-sample datasets. Nevertheless, from an overall perspective, the proposed method performs better on more datasets and obtains the highest average result. Furthermore, the W/T/L statistics reported in the table show that the proposed method maintains a clear overall advantage over the competing methods. In particular, it achieves a 9/0/1 record against DCSF, and also shows high win rates against CFR, JMI, MRMD, and mRMR. These findings indicate that, under the SVM classifier, the proposed method not only delivers strong average performance but also exhibits good stability and effectiveness across most datasets.

Table 3 presents the comparison results of average classification accuracy (mean ± std.) for all methods under the NB classifier. It can be seen that the proposed method achieves an average accuracy of 76.36% over the ten datasets, which still ranks first among all competing methods. It is slightly higher than DCSF (76.16%) and clearly superior to JMI (73.86%), MRMD (71.96%), mRMR (67.26%), and CFR (66.37%). These results indicate that the proposed method can also achieve strong overall classification performance under the naive Bayes classifier, showing particularly clear advantages over traditional mRMR and CFR methods.

Table 2. Average accuracy (mean±std.) with statistical significance on SVM

Datasets	CFR	DCSF	JMI	MRMD	mRMR	ARMR
Isolet	47.44±2.54	83.90±1.84	70.58±2.83	70.30±1.51	71.57±4.78	85.80±1.55
Lsvt	87.47±4.84	85.33±3.94	86.70±7.05	87.01±3.68	84.90±6.48	87.58±3.61
Movement_libras	60.40±7.28	78.34±6.54	73.71±8.40	76.74±7.33	76.71±9.13	79.55±5.64
WarpAR10P	65.43±14.83	68.24±12.84	82.17±12.61	81.89±9.09	81.35±13.50	83.54±12.81
WarpPIE10P	87.65±5.26	93.92±2.89	91.92±4.44	95.91±5.18	94.93±4.06	96.10±3.97
Yale_32x32	60.86±16.14	64.45±12.12	62.81±13.85	63.41±14.44	64.34±8.72	64.71±10.82
ALLAML	95.36±11.97	87.33±13.57	94.03±7.72	95.47±14.96	98.88±8.82	91.30±11.27
ColonX	85.48±11.36	81.88±14.45	86.16±10.21	85.18±13.40	84.52±13.38	82.49±11.04
Prostate-GE	93.67±6.71	94.33±6.39	94.24±7.48	93.24±5.35	91.72±7.36	89.10±8.89
TOX-171	68.30±12.00	82.06±6.47	70.18±13.79	74.07±12.80	72.03±8.42	85.66±7.81
Average	75.21	81.98	81.25	82.32	82.09	84.58
W/T/L	7/0/3	9/0/1	7/0/3	7/0/3	7/0/3	

Table 3. Average accuracy (mean±std.) with statistical significance on NB

ACC	CFR	DCSF	JMI	MRMD	MRMR	ARMR
Isolet	38.91±2.71	71.96±3.37	59.37±3.20	57.50±5.93	59.62±3.34	71.62±3.37
Lsvt	78.33±9.62	85.10±6.33	82.32±7.86	84.58±5.90	43.47±17.37	74.91±13.07
Movement_libras	45.54±7.50	62.95±9.83	56.01±11.65	59.05±9.44	58.10±6.59	64.97±8.73
WarpAR10P	60.12±14.42	61.87±15.11	70.17±10.87	63.67±9.09	63.62±9.47	69.49±9.28
WarpPIE10P	70.06±7.59	77.46±5.01	68.87±4.02	76.57±5.76	74.56±7.49	81.08±6.83
Yale_32x32	51.91±12.67	64.22±10.34	55.24±14.41	55.30±9.10	54.33±8.19	68.45±10.68
ALLAML	97.61±7.26	94.28±8.47	98.03±7.57	97.79±5.52	95.78±5.10	89.02±15.11
ColonX	74.60±16.78	82.01±13.24	82.95±8.10	73.08±18.87	74.71±21.06	83.20±13.67
Prostate-GE	93.84±7.71	92.23±8.13	94.91±5.25	90.63±3.52	90.43±4.73	85.04±9.17
TOX-171	52.82±7.19	69.55±15.14	70.76±6.93	61.43±8.34	57.96±12.71	75.81±11.95
Average	66.37	76.16	73.86	71.96	67.26	76.36
W/T/L	7/0/3	6/0/4	6/0/4	7/0/3	8/0/2	

From the results on individual datasets, the proposed method obtains the best performance on five datasets, namely Movement_libras, WarpPIE10P, Yale, ColonX, and TOX-171. In particular, on the TOX-171 dataset, the proposed method achieves an accuracy of 75.81%, which is noticeably higher than those of the competing methods. On the Yale and Movement_libras datasets, the proposed method also attains the best results, with accuracies of 68.45% and 64.97%, respectively, indicating that it has strong feature selection capability on some complex datasets. The proposed method also performs well on the WarpPIE10P and ColonX datasets, further demonstrating its good adaptability across different types of data.

However, compared with the results under the SVM classifier, the advantage of the proposed method under the NB classifier is less pronounced. For example, on the Isolet and Lsvt datasets, DCSF achieves higher classification accuracy, while on the ALLAML and Prostate-GE datasets, JMI or CFR performs better. This suggests that the naive Bayes classifier is more sensitive to assumptions about feature distributions, and the performance differences among feature selection methods under this classifier depend more strongly on the statistical characteristics of the data itself. Nevertheless, judging from both the overall average results and the W/T/L statistics, the proposed method still demonstrates strong competitiveness. Specifically, compared with CFR, MRMD, and mRMR, it achieves 7/0/3, 7/0/3, and 8/0/2, respectively; compared with DCSF and JMI, it obtains 6/0/4 and 6/0/4, respectively. These results show that, under the NB classifier, the proposed method still outperforms most competing algorithms overall.

Taken together, the results in Table 2 and Table 3 show that the proposed method achieves strong overall performance under both the SVM and NB classifiers. Its advantage is particularly evident under the SVM classifier, where it ranks first in terms of both average classification accuracy and the number of datasets on which it achieves the best result. This demonstrates that the proposed feature evaluation mechanism can effectively balance discriminative capability and redundancy suppression, thereby improving the quality of the selected feature subset. Meanwhile, although the advantage of the proposed method is somewhat less pronounced under the NB classifier on certain datasets, its overall average performance still surpasses that of the competing methods, indicating good classifier adaptability and a certain degree of robustness. Overall, the experimental results verify the effectiveness of the proposed method for feature selection in high-dimensional data classification tasks.

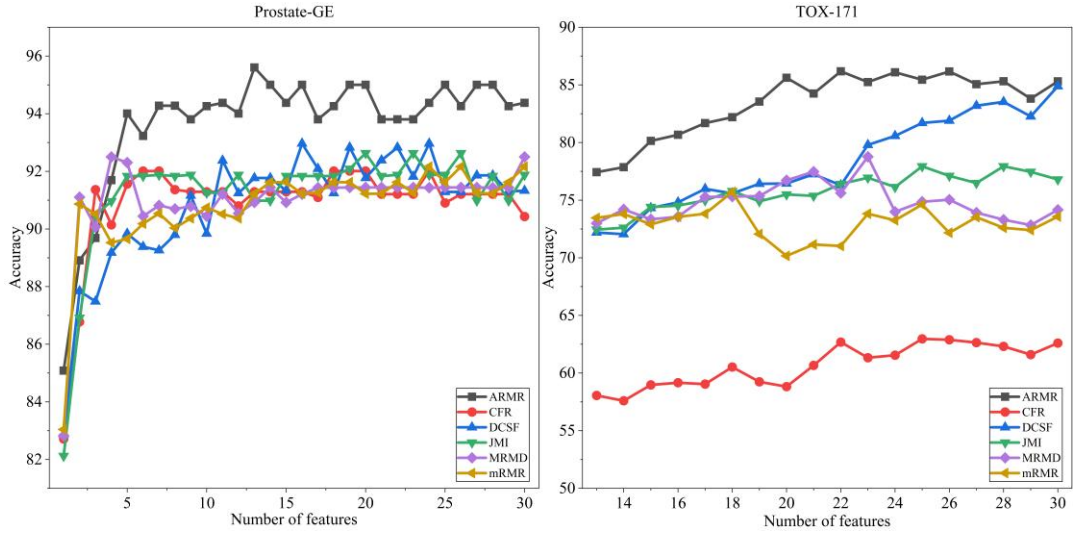

Figure 1. The average classification accuracy of SVM and NB on the dataset

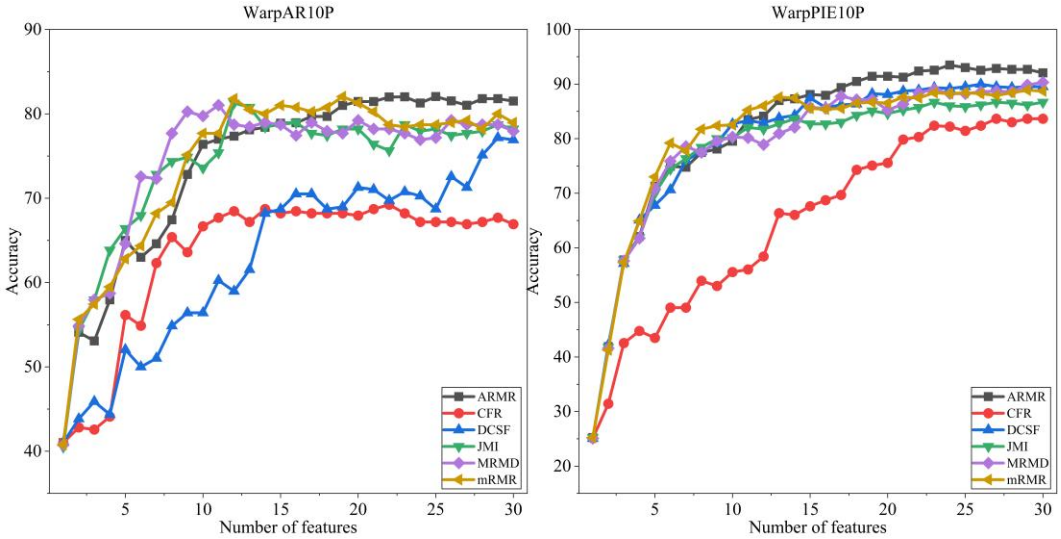

Figure 2. The average classification accuracy of SVM and NB on the dataset

To further illustrate the performance variation of ARMR during the feature selection process, Figure 1 and Figure 2 present the classification accuracy curves of ARMR and the other compared algorithms under different feature subset sizes. In this study, two representative medical datasets, Prostate-GE and TOX-171, as well as two datasets from other domains, WarpAR10P and WarpPIE10P, were selected for visualization. In the figures, the horizontal axis represents the size of the selected feature subset, while the vertical axis represents the average classification accuracy obtained by the SVM and NB classifiers.

As shown in Figure 1, on the Prostate-GE and TOX-171 datasets, the classification accuracy of all methods generally increases as the number of selected features grows and gradually becomes stable. ARMR maintains relatively high classification accuracy under most feature subset sizes, with its advantage being particularly evident on the TOX-171 dataset. This indicates that the proposed method is able to identify key features that contribute more effectively to the classification task at an early stage of the selection process.

As shown in Figure 2, on the WarpAR10P and WarpPIE10P datasets, the overall accuracy curve of ARMR remains among the top-performing methods and exhibits good stability as the number of selected features increases. In particular, on the WarpPIE10P dataset, ARMR achieves higher and more stable classification accuracy, indicating that the proposed method also has good adaptability to high-dimensional image data.

In addition to classification accuracy, F1-score and AUC are also important indicators for evaluating the performance of feature selection methods. Specifically, the F1-score provides a comprehensive measure of precision and recall, while AUC represents the area under the ROC curve. Both metrics range from 0 to 1, and values closer to 1 indicate better classification performance. Figure 3 and Figure 4 illustrate the distributions of F1-score and AUC for all algorithms over the ten datasets boxplots. In these plots, the lower and upper boundaries of each box correspond to the first quartile (Q1) and the third quartile (Q3), respectively, while the line inside the box represents the median and the red dot denotes the mean value. The height of the box, namely the interquartile range (IQR), reflects the dispersion of the middle 50% of the data; a smaller IQR indicates that the performance across different datasets is more concentrated and that the algorithm is more stable. The upper and lower whiskers represent the maximum and minimum values of the corresponding metric, respectively.

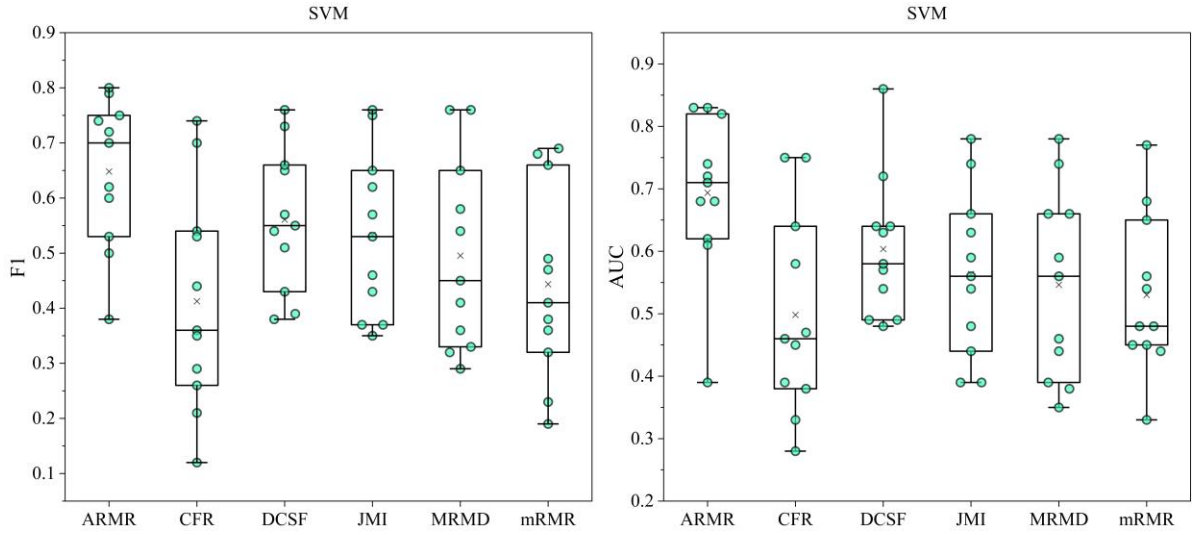

Figure 3. F1 and AUC metrics on SVM

Figure 3 shows the distributions of F1-score and AUC for all algorithms under the SVM classifier. It can be observed that ARMR achieves relatively favorable median values and overall distributions on both metrics, while its boxplots are relatively concentrated, indicating that the proposed method has good overall classification performance and stability across different datasets.

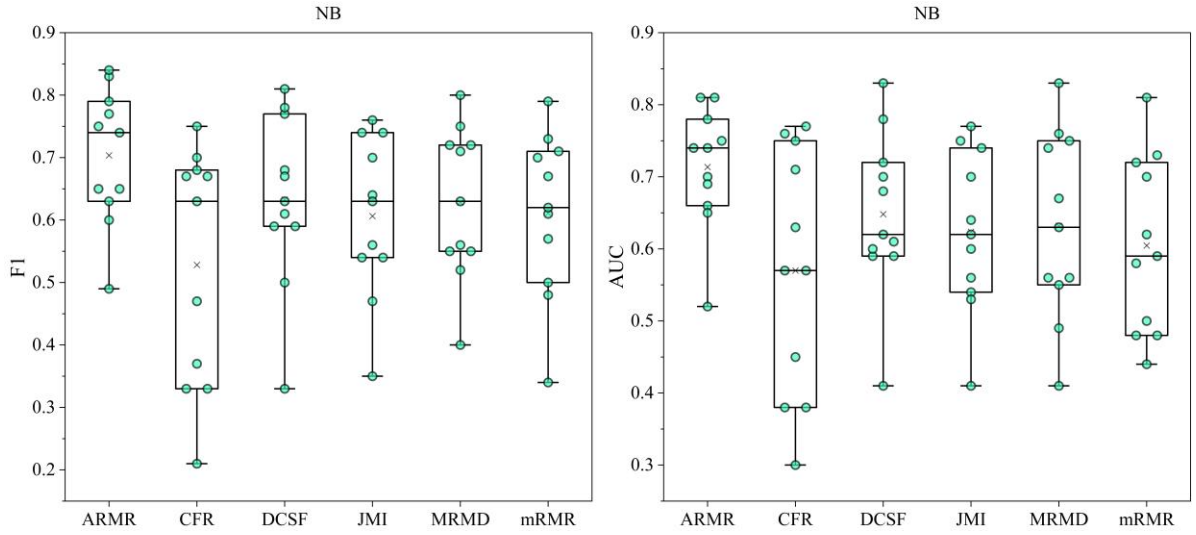

Figure 4. F1 and AUC metrics on NB

Figure 4 presents the distributions of F1-score and AUC for all algorithms under the NB classifier. Overall, ARMR still maintains favorable median values and relatively small ranges of variation on both metrics, further demonstrating that the proposed method possesses a certain degree of robustness under different classifiers.

5. Conclusion

This paper focuses on the problem of filter-based feature selection for high-dimensional medical data. To address the limitation of traditional information-theoretic methods, in which the redundancy penalty is usually fixed and cannot be dynamically adjusted according to the actual relationships among features, a filter-based feature selection method, termed ARMR, is proposed based on joint mutual information and an adaptive redundancy penalty. Specifically, joint mutual information is employed to characterize the joint discriminative ability of candidate features and selected features with respect to the class label, while an adaptive penalty factor is introduced to dynamically adjust the strength of redundancy suppression. In this way, a comprehensive evaluation function is constructed that balances feature complementarity and redundancy suppression. On this basis, a complete methodological framework is established, including the feature selection procedure, algorithm pseudocode, and time complexity analysis.

In the experimental part, the proposed method was evaluated on 10 datasets of different types and compared with CFR, DCSF, JMI, MRMD, and mRMR. The results show that ARMR achieves strong overall performance under both the SVM and NB classifiers. In particular, under the SVM classifier, ARMR exhibits a more significant advantage, ranking first in both average classification accuracy and the number of datasets on which it achieves the best results. Under the NB classifier, ARMR also maintains strong competitiveness. Further analyses based on classification accuracy curves under different feature subset sizes, as well as boxplots of F1-score and AUC, demonstrate that the proposed method not only effectively improves classification performance but also exhibits good stability and a certain degree of robustness. Overall, the proposed method shows good effectiveness and promising application potential in feature selection tasks for high-dimensional medical data.

In future work, the proposed method can be further

improved in several aspects. First, more robust discretization and probability estimation strategies will be considered to further enhance the reliability of information measures in small-sample scenarios. Second, the proposed evaluation mechanism will be extended to imbalanced data, multi-label data, and multimodal medical data, so as to broaden the applicability of the method. Third, the proposed method will be combined with embedded learning or intelligent optimization techniques to further improve the quality of the selected feature subset while maintaining computational efficiency.

References

[1] Y. Kang, D. Zheng, H. Wang, Y. Peng, and S. Zhou, 'Dual-metric guided multi-strategy hybrid optimization for feature selection on high-dimensional medical data', Swarm Evol. Comput., vol. 98, p. 102118, Oct. 2025, doi: 10.1016/j.swevo.2025.102118.

[2] M. Wang, A. A. Heidari, and H. Chen, 'A multi-objective evolutionary algorithm with decomposition and the information feedback for high-dimensional medical data', Appl. Soft Comput., vol. 136, p. 110102, 2023, doi: https://doi.org/10.1016/j.asoc.2023.110102.

[3] D. Dhinakaran, L. Srinivasan, S. E. Raja, K. Valarmathi, and M. G. Nayagam, 'Synergistic feature selection and distributed classification framework for high-dimensional medical data analysis', MethodsX, vol. 14, p. 103219, 2025, doi: https://doi.org/10.1016/j.mex.2025.103219.

[4] X. Zhang and J. Wang, 'A noise-robust feature selection using KNN and weighted fuzzy rough sets for imbalanced multi-scale data', Appl. Soft Comput., vol. 194, p. 114964, 2026, doi: https://doi.org/10.1016/j.asoc.2026.114964.

[5] H. Ju et al., 'Distributed multi-label feature selection via feature-label information granulation', Inf. Sci., vol. 742, p. 123334, 2026, doi: https://doi.org/10.1016/j.ins.2026.123334.

[6] F. Deng, L. Zhao, N. Yu, Y. Lin, and L. Zhang, 'Union With Recursive Feature Elimination: A Feature Selection Framework to Improve the Classification Performance of Multicategory Causes of Death in Colorectal Cancer', Lab. Invest., vol. 104, no. 3, p. 100320, 2024, doi: https://doi.org/10.1016/j.labinv.2023.100320.

[7] S.-W. Lin, Z.-J. Lee, S.-C. Chen, and T.-Y. Tseng, 'Parameter determination of support vector machine and feature selection using simulated annealing approach', Appl. Soft Comput., vol. 8, no. 4, pp. 1505–1512, 2008, doi: https://doi.org/10.1016/j.asoc.2007.10.012.

[8] L. Gao, L. Li, and X. Chen, 'Sparse learning of interval type-2 fuzzy model with embedded feature and rule selection', Neurocomputing, vol. 677, p. 133102, 2026, doi: https://doi.org/10.1016/j.neucom.2026.133102.

[9] T. Dupré la Tour, M. Eickenberg, A. O. Nunez-Elizalde, and J. L. Gallant, 'Feature-space selection with banded ridge regression', NeuroImage, vol. 264, p. 119728, Dec. 2022, doi: 10.1016/j.neuroimage.2022.119728.

[10] H. Lyu, M. Wan, J. Han, R. Liu, and C. Wang, 'A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining', Comput. Biol. Med., vol. 89, pp. 264–274, 2017, doi: https://doi.org/10.1016/j.compbiomed.2017.08.021.

[11] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. Wang, 'A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information', Eng. Appl. Artif. Intell., vol. 131, p. 107865, May 2024, doi: 10.1016/j.engappai.2024.107865.

[12] U. Agrawal, V. Rohatgi, and R. Katarya, 'Normalized Mutual Information-based equilibrium optimizer with chaotic maps for wrapper-filter feature selection', Expert Syst. Appl., vol. 207, p. 118107, 2022, doi: https://doi.org/10.1016/j.eswa.2022.118107.

[13] W. Gao, L. Hu, P. Zhang, and J. He, 'Feature selection considering the composition of feature relevancy', Pattern Recognit. Lett., vol. 112, pp. 70–74, Sep. 2018, doi: 10.1016/j.patrec.2018.06.005.

[14] W. Gao, L. Hu, and P. Zhang, 'Class-specific mutual information variation for feature selection', Pattern Recognit., vol. 79, pp. 328–339, Jul. 2018, doi: 10.1016/j.patcog.2018.02.020.

[15] 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy'. Accessed: Dec. 30, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/1453511

[16] 'Data visualization and feature selection | Proceedings of the 12th International Conference on Neural Information Processing Systems'. Accessed: Dec. 30, 2024. [Online]. Available: https://dl.acm.org/doi/10.5555/3009657.3009755

[17] M. Bennasar, Y. Hicks, and R. Setchi, 'Feature selection using Joint Mutual Information Maximisation', Expert Syst. Appl., vol. 42, no. 22, pp. 8520–8532, Dec. 2015, doi: 10.1016/j.eswa.2015.07.007.

[18] Z. Zeng, H. Zhang, R. Zhang, and C. Yin, 'A novel feature selection method considering feature interaction', Pattern Recognit., vol. 48, no. 8, pp. 2656–2666, Aug. 2015, doi: 10.1016/j.patcog.2015.02.025.

[19] W. Gao, L. Hu, and P. Zhang, 'Feature redundancy term variation for mutual information-based feature selection', Appl. Intell., vol. 50, no. 4, pp. 1272–1288, Apr. 2020, doi: 10.1007/s10489-019-01597-z.

[20] 'Feature Selection: A Data Perspective: ACM Computing Surveys: Vol 50, No 6'. Accessed: Dec. 30, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3136625

[21] Dias Daniel, Peres, Sarajane and H. Bscaro, 'Libras Movement'. 2009.

[22] A. Tsanas, 'LSVT Voice Rehabilitation'. 2014.