Complex & Intelligent Systems (2023) 9:6139–6153
https://doi.org/10.1007/s40747-023-01039-x

ORIGINAL ARTICLE

Feature selection using non-dominant features-guided search for gene expression profile data

Xiaoying Pan[1] · Jun Sun[2] · Huimin Yu[2] · Yufeng Xue[2]

Received: 1 August 2021 / Accepted: 9 March 2023 / Published online: 26 April 2023

Abstract
Gene expression profile data have high-dimensionality with a small number of samples. These data characteristics lead to a long training time and low performance in predictive model construction. To address this issue, the paper proposes a feature selection algorithm using non-dominant feature-guide search. The algorithm adopts a filtering framework based on feature sorting and search strategy to overcome the problems of long training time and poor performance. First, the feature pre-selection is completed according to the calculated feature category correlation. Second, a multi-objective optimization feature selection model is constructed. Non-dominant features are defined according to the Pareto dominance theory. Combined with the bidirectional search strategy, the Pareto dominance features under the current category maximum relevance feature are removed one by one. Finally, the optimal feature subset with maximum correlation and minimum redundancy is obtained. Experimental results on six gene expression data sets show that the algorithm is much better than Fisher score, maximum information coefficient, composition of feature relevancy, mini-batch K-means normalized mutual information feature inclusion, and max-Relevance and Min-Redundancy algorithms. Compared to feature selection method based on maximum information coefficient and approximate Markov blanket, the algorithm not only has high computational efficiency but also can obtain better classification capabilities in a smaller dimension.

Keywords High dimensional and small-sample size · Feature selection · Distance measurement · Pareto dominance theory · Maximum correlation and minimum redundancy

Introduction

With bio-information development, gene expression profile analysis has become an essential means of oncogene identification and plays a critical role in cancer classification and prediction. Microarray technology [1] uses many probes each time, and gene information involves many aspects, leading to the high-dimensionality of microarray data. At the same time, the sample preparation cost is high, and the process is complex, which leads to the small-sample size and the uneven distribution of sample categories. Therefore, gene expression profile data are typical high-dimensional small-sample data [2], which has strong feature redundancy and all the characteristics of high-dimensional small-sample data. This type of data is directly used to build predictive models, and it is easy to have problems such as long training time, low model performance, and overfitting. Feature selection is required to eliminate and mitigate dimensional disasters, improve model performance, reduce runtime, and extract beneficial information [3].

Literature [4] summarizes the popular feature selection methods broadly divided into the filter, wrapper, and embedded methods [5]. Wrappers select subsets of features from the initial feature collection, train learners such as support vector machine (SVM) classifier, and evaluate subsets based on the

✉ Jun Sun
15389096817@163.com

Xiaoying Pan
xiaoying_pan@163.com

Huimin Yu
1500314260@qq.com

Yufeng Xue
1337819533@qq.com

[1] Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China

[2] School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China

learner's performance. Wrapper performs well in classification, but it costs too much and risks overfitting. Embedded methods automatically select features during training. This kind of method, although short in computing time, relies too much on classifiers. Filtering methods can be scaled efficiently on high-dimensional datasets, regardless of classifiers, and are particularly widely used in high-dimensional data.

In high-dimensional data, the removal of redundancy is a hot topic of research [6]. Filtering methods for feature redundancy problems based on information metrics [7], such as maximum relevancy and minimum redundancy feature selection (mRMR) [8], fast correlation-based filter (FCBF) [9], markov blanket based feature selection algorithm [10], and related improvement algorithms promoted based on the above algorithms [11]. These methods are not suitable for high-dimensional data processing because of large computation and high time complexity.

The paper proposes a feature selection algorithm using a non-dominant features-guided search (NDFS) to solve the above problems. The main ideas of this method are as follows: (1) Based on the framework combining feature ranking and search strategy, the irrelevant, redundant features can be quickly filtered and screened. (2) Fisher score and cosine distance measure the class correlation and similarity of features. (3) The concept of non-dominant features is proposed, and a two-way search strategy is adopted in the process of non-dominant feature-guided search. (4) Finally, a feature subset with maximum correlation and minimum redundancy is selected to improve the performance of subsequent classification.

The rest of this paper is organized as follows. "Related work" discusses related work. "Preliminaries" presents some preliminaries for this work. A novel feature selection method is given in "Proposed architecture and methods". "Experimental results and discussion" gives experimental results and discussion. "Conclusion" concludes the paper.

Related work

This section first briefly introduces the nature of the feature selection problem, and then discusses the existing mainstream feature selection approaches.

Feature selection problem

Mathematically, the feature selection problem can be expressed in the following way. Assuming a dataset S contains d features. The essence of the feature selection problem is to select relevant features among d features, to optimize the given classification performance index as much as possible. Given a dataset $S = \{f_1, f_2, f_3, \text{----}, f_d\}$, the goal is to select the optimal subset of features from S. Select a subset $D = \{f_1, f_2, f_3, \cdots, f_n\}$, where $n < d, f_1, f_2, f_3, \cdots, f_n$ represent the features of the dataset.

Existing feature selection approaches

Feature selection plays a critical role in classification problems, especially for data sets that have many features [12]. These features need to be measured in two ways: correlation between features and class and redundancy between features. Combined with the corresponding search strategy, the final feature subset is obtained [13].

Early feature selection methods only consider selecting features more relevant to categories. Relevant features can be derived from label information. The ranking of features by scoring them based on relevancy criterion, represented by Relief [14], ReliefF [15], Fisher score [16] and Maximal Information Coefficient (MIC) [17]. The relief method and its multi-class extension, ReliefF, select features from instances that are separated from different classes. The algorithm randomly selects an instance from the data, then calculates the distance to find positive or negative samples of its nearest neighbor, and updates the weight of each feature. The Relief series algorithms operate efficiently with no restrictions on data types. Fisher score algorithm uses probability distance as the evaluation criterion of a feature. The distance between the same class of samples is small, and the distance between different classes of samples is large. The Fisher score algorithm is versatile, has low time complexity, and is particularly suitable for working with high-dimensional datasets. However, in feature selection, there are many redundancy features because these algorithms fail to consider the relationship between features and features.

Redundant features do not provide any additional information other than noise for the classification algorithm, so they should be removed. Typical algorithms include mRMR, correlation-based feature selection (CFS) [18], composition of feature relevancy (CFR) [19]. mRMR is based on mutual information, minimizing the correlation between features and mutual information, and maximizing the correlation between features and class labels. In addition, many literatures are modified based on mRMR method, such as using normalized mutual information [20] and various monotonic dependence measures to replace mutual information for feature selection. CFS is evaluated based on the predictive power of each feature in the subset and its correlation, and the subsets of individual features with strong predictive power and low correlation within the feature subset perform well. CFR calculates the relevance score by calculating union condition information for candidate characteristics for a given selected feature collection category. Feature redundancy is given by the joint information of candidate features, categories, and selected features.

The calculation of feature redundancy is highly complex on high-dimensional data. Scholars adopt a two-stage feature selection method to balance correlation and redundancy to improve efficiency [21]. The basic idea is to calculate the correlation between features and categories for sorting and then use the algorithm based on a search strategy to remove redundant features. The typical algorithm is FCBF proposed in 2004. Firstly, it calculates the symmetrical uncertainty (SU) of each feature and class and sorts it in descending order, removing features less than pre-set thresholds, i.e., irrelevant features. Secondly, this algorithm selects the feature with the largest SU of the feature and class in the current feature set. It calculates the SU of the remaining features and the current features and the SU of the remaining features and class one by one until all redundant features under the feature are removed. Feature Selection Method Based on maximum information coefficient and approximate markov blanket (FCBF-MIC) [22] still uses symmetric uncertainty to measure the correlation between features and categories in the first stage. In the second stage, an approximate Markov blanket is used to fuse the maximum information coefficient measurement standard to remove redundant features. Feature selection algorithm based on approximate Markov blanket is proposed and named as normal max-relevance and min-redundancy (nmRMR) [23] algorithm. Firstly, the features are sorted using the maximum correlation minimum redundancy criterion. Secondly, the irrelevant and redundant features are removed according to the approximate Markov blanket condition. Mini batch K-means normalized mutual information feature inclusion (KNFI) [24] is proposed in 2019, which combines filter and wrapper techniques. The algorithm uses normalized mutual information as a measure to sort the features after clustering by small-batch K-means, the sorting features of the first stage are added to the subset one by one.

The above method can select the relevant features and eliminate the redundant features. However, they still need to be improved in determining the optimal subset of features efficiently and improving the classification performance.

Preliminaries

Fisher score has been maturely applied to feature selection problems. Pareto dominance theory is mainly used to deal with multi-objective optimization problems and the essence of feature selection problems is a multi-objective problem. This section introduces Fisher score algorithm and Pareto dominance theory.

Fisher score algorithm based on probability distance standard

Fisher score is a correlation-based feature evaluation criterion based on probability distance, a practical feature selection method. In the Fisher score algorithm, intra-class dispersion S_w represents intra-class distance, and inter-class dispersionS_b represents inter-class distance. The class distinguishing ability of a feature is the ratio of S_b to S_w. The larger this value is, the stronger the category correlation of the feature is.

Assuming there is a binary classification problem. The positive sample is marked as 1, the negative sample is marked as 0, the number of positive samples is n_1, the number of negative samples is n_0, the total number of samples is n, and the number of features is m.

S_w, S_b of feature f are defined as for Eqs. (1) and (2).

$$S_w^{(f)} = n_0\left(\sigma_0^{(f)}\right)^2 + n_1\left(\sigma_1^{(f)}\right)^2 \tag{1}$$

$$S_b^{(f)} = n_0\left(\mu_0^{(f)} - \mu^{(f)}\right)^2 + n_1\left(\mu_1^{(f)} - \mu^{(f)}\right)^2 \tag{2}$$

The correlation between feature f and class calculated by Fisher score is defined as Eq. (3).

$$Correlation\,(f) = FS(f) = \frac{S_b^{(f)}}{S_w^{(f)}} \tag{3}$$

where $\mu_0^{(f)}$ is the mean of feature f in the negative sample, $\mu_1^{(f)}$ is the mean of feature f in the positive sample, and $\mu^{(f)}$ is the mean of feature f in the overall sample. $\sigma_0^{(f)}$ and $\sigma_1^{(f)}$ are the variance of feature f in negative and positive samples.

It can be seen from formula (3) that the greater the inter-class dispersion of a feature, the smaller the intra-class dispersion, and the better the classification effect of the feature.

Pareto dominance theory

Multi-objective optimization generally involves maximizing or minimizing multiple objective functions. Generally, minimizing a multi-objective optimization function [25] can be described as a formula (4).

$$\begin{aligned} &\min\ j(x) = (j_1(x),\ j_2(x),\ \ldots,\ j_k(x))^T \\ &s.t.\quad g_i(x) \geq 0,\ i = 1,\ 2,\ \cdots m; \\ &\qquad h_j(x) = 0,\ j = 1,\ 2,\ \cdots,\ l \end{aligned} \tag{4}$$

In formula (4), x is the decision variable $j_1(x)$, $j_2(x)$, ..., $j_k(x)$ represent k objective functions, and the objective is to minimize them. $g_i(x)$, $h_j(x)$ is the constraint condition of the problem.

In this minimized multi-objective optimization problem, for k objective components, given any two decision variables x_a and x_b. If the following two conditions are true, then x_a dominates x_b·[26].

$$(1)\ \forall i \in 1, 2, \ldots, k, \quad j_i(x_a) \leq j_i(x_b)$$
$$(2)\ \exists i \in 1, 2, \ldots, k, \quad j_i(x_a) < j_i(x_b)$$

When the value of the objective function corresponding to the solution x_a is better than the value of the objective function corresponding to the solution x_b, x_a is called strong Pareto dominating x_b. When a solution x_a, there is no other solution that can dominate it, then it is called non-dominated solution.

Proposed architecture and methods

In this paper, a feature selection using non-dominant features-guided search is proposed. The Fisher score algorithm, based on the probability distance standard, measures the category correlation of features and extracts a set of pre-selected features with high correlation. Cosine similarity is used to measure the similarity between features based on geometric distance measurement standards, and sample features with lower dimensions represent more information of samples. The Pareto dominance theory is introduced to calculate the non-dominant features for guided search. The feature subset with the most significant category correlation and the least redundancy between features is selected. Figure 1 shows the overall architecture of this method.

Cosine similarity measure

Cosine similarity [27] is a method to measure the similarity of two vectors. This paper introduces it to calculate the similarity between features. Assuming there are two features f_1 and f_2, where they represent any two features from a feature set, the cosine value of the two features can be used to measure the similarity between the two features. The cosine similarity of the two features f_1 and f_2 was calculated using Eq. (5).

$$Similarity\ Info(f_1,\ f_2) = cos(\theta) = \frac{f_1 . f_2}{\|f_1\| \times \|f_2\|} = \frac{\sum_{i=1}^{n} (f_{1_i} f_{2_i})}{\sqrt{\sum_{i=1}^{n} f_{1_i}^2} \times \sqrt{\sum_{i=1}^{n} f_{2_i}^2}} \quad (5)$$

In Formula (5), given f_1 and f_2 as the feature vectors, n as the total number of instances. f_{1_i} and f_{2_i} represent the values of f_1 and f_2 corresponding to the ith instance, respectively.

Therefore, the range of feature similarity is $[-1, 1]$. The closer the value is to 1, the more similar the two features are.

Pareto dominance feature

In the process of feature selection, not only the correlation between features and class but also the correlation between features should be considered. Features with high-class correlation may be similar in data distribution, so the similarity between features and features may be high. It means there may be redundancy between features. High redundancy can not improve the model's performance and even make the performance of the model decline sharply, so it is necessary to remove redundancy. Thus, the feature selection process can regard it as a multi-objective optimization problem. The goal is to select features with the highest class correlation and the lowest feature substitutability. Inspired by the Pareto theory in the multi-objective issue, the concept of the Pareto dominance feature is defined

$$\begin{cases} f_1 > f_2 (f_1 \text{dominates } f_2), & if\, C(f_1) > C(f_2) \text{and } \mathrm{S}(f_1,\ f_2) > \mu \\ f_1 \geq f_2 (f_1 \text{weakly dominates} f_2), & else if\, C(f_1) \geq C(f_2) \text{and } \mathrm{S}(f_1,\ f_2) > \mu \\ f_1 \sim f_2 (f_1 \text{ non} - \text{dominates} f_2), & else \end{cases} \quad (6)$$

Assuming there are two features f_1 and f_2, if the category correlation of f_1 is higher than that off_2, and the similarity between f_1 and f_2 is higher than the given threshold, then f_1 dominancef_2. Otherwise, f_1 is a non-dominated feature of f_2.

Definition 1 Non-dominance feature. For any two features, f_1 and f_2, the binary relationship $>$, $\geq$ and $\sim$ are defined as for formula (6).

Where, given $C(f_i)$ as the category correlation of feature f_i. $C(f_1) > C(f_2)$ represents that the category correlation of f_1 is higher than that of f_2.$\mathrm{S}(f_1,\ f_2)$ represents feature similarity between f_1 and f_2.μ is the given similarity threshold.

In the NDFS algorithm proposed in this paper, the category correlation of features is calculated by Fisher score, and the similarity between features is calculated by cosine similarity.

The algorithm procedure

The specific process of the algorithm is shown in Table 1.

Assuming the original dataset has n samples, m genes, and the feature set $F = \{f_1,\ f_2, \cdots, f_m\}$. Assuming the threshold set in Fisher score algorithm is μ_1, and the similarity threshold is μ_2 when eliminating redundant features based on the non-dominated theory. Proposed algorithm proceeds as follows.

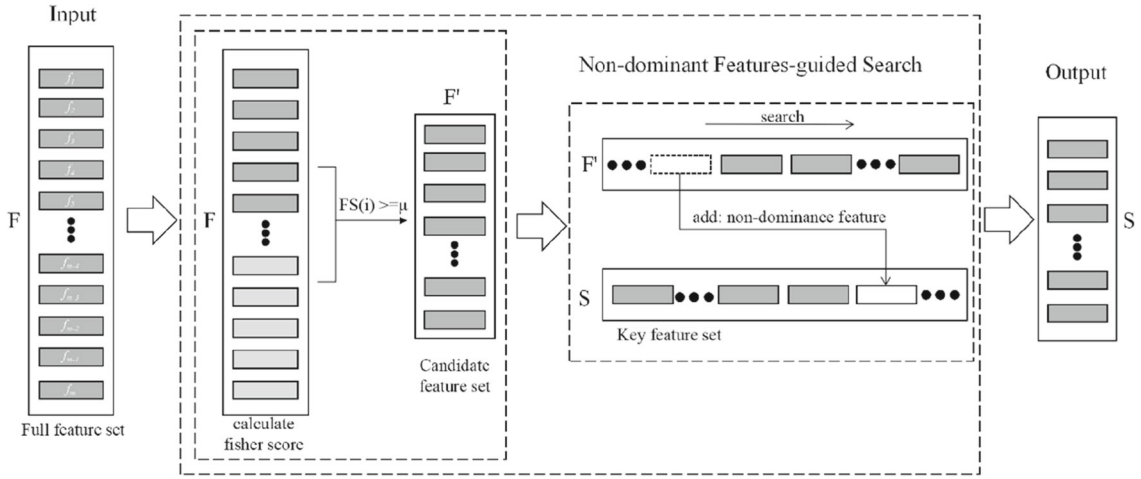

Fig. 1 NDFS architecture

Step 1: Fisher score of each feature is calculated using the Fisher score algorithm. The features with small class correlation are removed by thresholds μ_1 to obtain a subset F of the remaining features.

Step 2: Given an empty set S and selects the first feature f_k from the remaining feature subset F that gives the largest Fisher score between the feature and the class target Y.
The feature f_k is added to the selected key feature set S (i.e., $S \leftarrow S \cup f_k$) and then removed from the set F (i.e., $F \leftarrow F \backslash f_k$).

Step 3: Select the largest Fisher score feature f_d in the remaining feature subset, where the class correlation of feature f_s in the key feature subset is greater than that of feature f_d (i.e., $C(f_s) > C(f_d)$). The similarity between f_s and f_d is calculated. If $S(f_s, f_d) > \mu_2$, the feature f_d is dominated by f_s, and the feature f_d is removed from the remaining set. If $S(f_s, f_d) \leq \mu_2$, the feature f_d is not dominated by f_s. If the key feature subset does not have the dominated feature f_d, it is added to the key feature subset.

Step4: If the remaining set F is empty, terminate the algorithm. Otherwise, go to step 3. The final output is the key feature subset S containing non-dominated features, which is the feature subset with the maximum correlation and minimum redundancy.

Experimental results and discussion

This section uses the proposed algorithm to discuss the experimental results on six high-dimensional small-sample data sets. Section "Experimental data sets" introduces the experimental data sets in detail. Section "Experimental setup" gives the experimental setup. Section "Feature selection experiment" describes the feature selection process experiment of NDFS in detail. And the feature selection results of the proposed method are compared with 6 algorithms in Section "Experimental result analysis".

Experimental data sets

To verify the effectiveness and applicability of this method in dealing with the feature selection problem of high-dimensional and small-sample gene data, this paper selects six public gene data sets. The HeadNeck data set is obtained from the GEO database [28], the two public data sets Colon data set and the Leukemia data set are obtained from Kaggle, the Lung dataset, and the 11_Tumors dataset are obtained from the website http://www.gemssystem.org/, and the LIHC dataset is from the cancer genome atlas (TCGA). These data have been widely cited by scholars at home and abroad and have certain standards. Table 2 summarizes the 6 public datasets used in this study. It contains 4 binary datasets and 2 multi-category datasets.

The Colon dataset consists of 62 samples collected from Colon cancer patients, including 40 tumor samples and 22 normal samples. The Leukemia dataset contains 72 case samples of 2 different leukemias, including acute myeloid (AML) and acute lymphoblastic leukemia (ALL). The 11_Tumors dataset contains 174 samples and genetic data of 11 common human cancer cases, including prostate cancer, bladder/urethral cancer transitional cell carcinoma and squamous cell carcinoma) invasive breast ductal carcinoma, rectal cancer, gastric adenocarcinoma, clear kidney cell carcinoma, liver cancer, ovarian serous papillary adenocarcinoma, pancreatic cancer and Lung adenocarcinoma and squamous cell carcinoma). The Lung dataset contains four different Lung tumors (139 cases of adenocarcinoma, 6 cases of small cell Lung cancer, 21 cases of squamous cell carcinoma, and 20

Table 1 Pseudocode of NDFS

Input: F: original dataset, Y: class label, μ_1, μ_2: threshold

Process:

1. **for** each feature $f_i \in F$
2. $FisherScore(i) = FS(f_i)$
3. **if** $FisherScore(i) < \mu_1$
4. $F = F \setminus \{f_i\}$
5. **end if**
6. **end for**
7. $f_k = \underset{f_i \in F}{\operatorname{argmax}}\left(FisherScore(f_i)\right)$
8. $S = \emptyset$ (empty set)
9. $S = S \cup \{f_k\}$
10. $F = F \setminus \{f_k\}$
11. **While** $F \neq \emptyset$
12. $f_d = \underset{f_j \in F}{\operatorname{argmax}}\left(FisherScore(f_j)\right)$
13. $i = 0$
14. **for** each feature $f_s \in S$
15. $i = i + 1$
16. **if** $SimilarityInfo(f_s, f_d) > \mu_2$
17. $F = F \setminus \{f_d\}$
18. **break**
19. **else**
20. **if** $i == num(S)$
21. $S = S \cup \{f_d\}$
22. $F = F \setminus \{f_d\}$
23. **continue**
24. **end if**
25. **end if**
26. **end for**
27. **end while**

Output: Key feature subset S

Table 2 Datasets used in experiments

Data set	#Features	#Samples	#Classes	Class distribution
Colon	2000	62	2	(22/40)
Leukemia	7129	72	2	(57/35)
11_Tumors	12,533	174	11	(27/8/26/23/12/11/7/26/6/14/14)
Lung	12,600	203	5	(139/17/6/21/20)
HeadNeck	12,727	105	2	(50/55)
LIHC	54,869	424	2	(50/374)

Table 3 Parameter table of each model

Model	Hyperparameter
Support vector machine (SVM)	($C = 1.0$, kernel = ''rbf'')
Decision tree (DT)	(Criterion = 'entropy', max_depth = 20)
Random forest (RF)	(n_estimators = 200, criterion = 'entropy', max_depth = 4)
Logic regression (LR)	($C = 1.0$, penalty = 'l2')
Multi-layer perception machine (MLP)	(Activation = 'relu', hidden_layer_sizes = (100,), solver = 'adam')

cases of Lung carcinoid) and 17 cases of normal Lung tissue. The HeadNeck dataset contains 55 samples with local recurrence and 50 samples without local recurrence. The LIHC dataset from TCGA includes 374 liver cancer samples and 50 paracancerous samples. The number of samples in these datasets ranges from 62 to 424, and the number of features ranges from 2000 to 54,869, all of which are high-dimensional small-sample data.

Experimental setup

To evaluate the classification performance of the selected feature set, SVM, decision tree (DT), random forest (RF), logic regression (LR), and multi-layer perception machine (MLP) are selected to construct prediction models. The parameter settings of each model are shown in Table 3. AUC values, Accuracy, F1-score, and ROC are used as evaluation indicators to evaluate the performance of different feature results and the constructed prediction model.

To avoid overfitting and improve data reusability, fivefold cross-validation is used in this experiment. The original data set is randomly divided into five equal parts. The proportion of positive and negative samples is consistent with the original data's proportion of positive and negative samples in each equal part. One sample is selected as the test set, and the other four samples are selected as the training set. Five models are finally obtained after five times of execution. The evaluation indexes of these five models are taken as the final prediction model's evaluation results by calculating the average value.

For performance comparison, the following algorithms have been selected: Fisher score, MIC, FCBF-MIC, CFR, KNFI, and mRMR. Since the Fisher score, MIC, CFR, and mRMR algorithms can directly select a certain number of features, the number of features selected by these two methods is consistent with the number of features finally selected by NDFS. Since the number of features that FCBF-MIC, and KNFI eventually generate cannot be known in advance, there is no limit to the number of features that the method ultimately selects. FCBF-MIC is directly consistent with the NDFS during the pre-selection process.

Feature selection experiment

When using the NDFS algorithm for feature selection, two thresholds need to be determined. One is the Fisher score threshold μ_1 for pre-selection using the Fisher score algorithm, and the other is the similarity threshold μ_2 for removing redundant features based on Pareto dominant theory.

To determine the threshold μ_1, the original features are sorted according to the Fisher score value. For visual observation, the top 50, 100, 200, 300, 400, 500 features are selected to form a series of feature subsets. According to these feature subsets, SVM and logistic regression (LR) is used to construct the classification model. The average classification accuracy is used to evaluate the feature subset. Finally, the Fisher score value corresponding to the appropriate number of features is selected to determine the threshold of each data set. Finally, the corresponding Fisher score is selected based on the appropriate number of features to determine each dataset's threshold μ_1.

Figure 2 shows the average classification accuracy corresponding to different features on the 6 datasets. It can be seen from Fig. 2a that for Colon dataset, when the number of features is about 400, the accuracy of the SVM and LR classifiers is relatively high. At this time, the corresponding fisher value is 0.0627, so the threshold μ_1 of Colon dataset is determined to be 0.06, and the final number of selected features is 414. It can be seen from Fig. 2b that for the Leukemia dataset, when the number of features is about 400, the accuracy of the SVM and LR classifiers is relatively high. Therefore, the threshold μ_1 corresponding to the Leukemia dataset is determined to be 0.2, and the number of selected features is 406. As can be seen from Fig. 2c, for the 11_Tumors dataset, when the number of features is about 300, the accuracy of the classifier is relatively high. Therefore, the threshold μ_1 corresponding to the 11_Tumors dataset is determined to be 1.13, and the number of selected features is 302. It can be seen from Fig. 2d that for the Lung dataset, when the number of features is about 300, the accuracy of the classifier is relatively high. Therefore, the threshold μ_1 corresponding to the Lung dataset is determined to be 1.17, and the number of selected features is 299. It can be seen from Fig. 2e that for the HeadNeck dataset, when the number of features is about 300, it performs best on the LR classifier. Although the accuracy of SVM classifier increase with the decrease of the number of features, the performance of the LR classifier decreases gradually. To avoid losing important features, the number of features is about 300. Therefore, the threshold μ_1 corresponding to the HeadNeck dataset is determined to be 0.05, and the number of selected features is 287. It can be

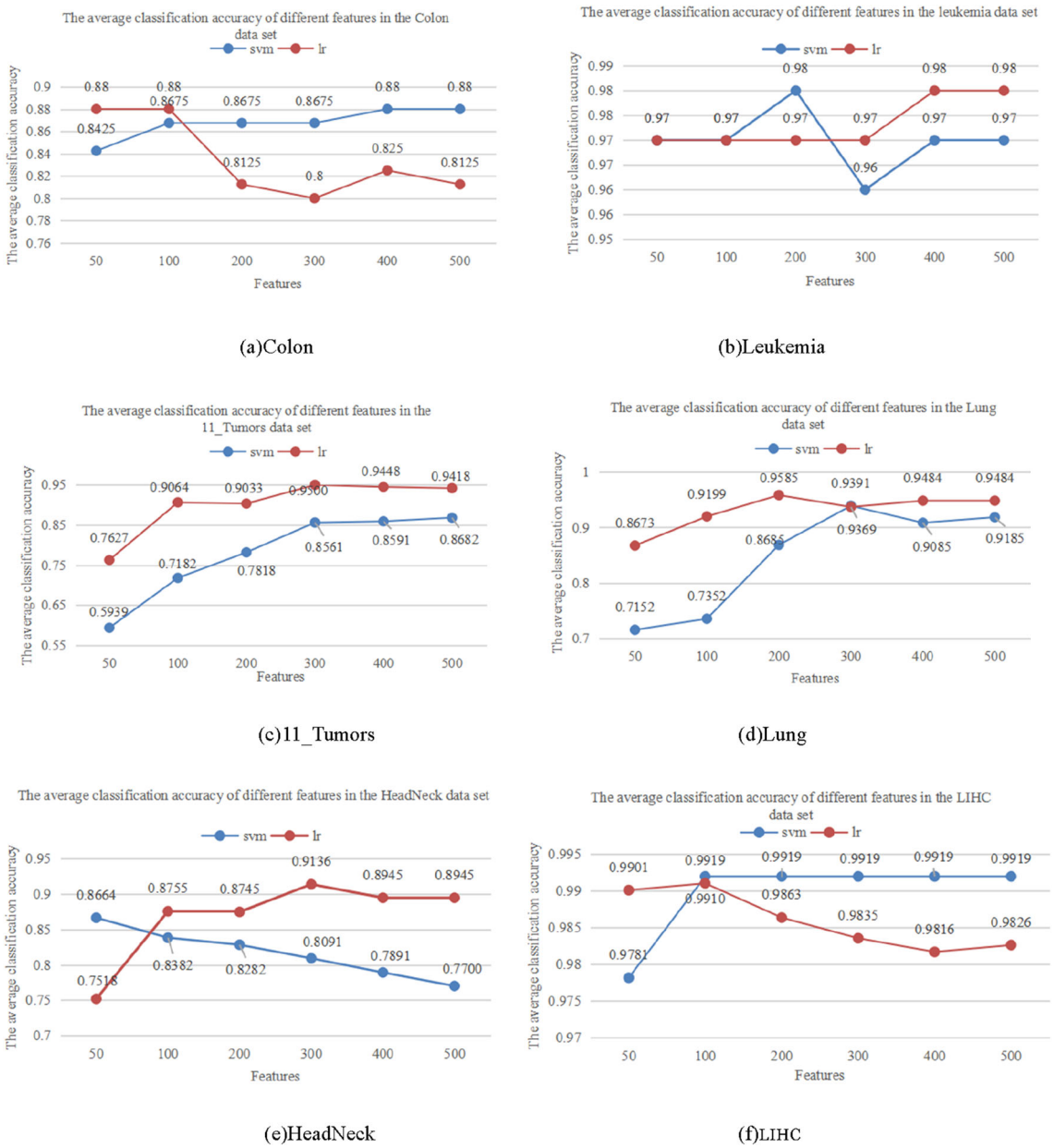

Fig. 2 Average classification accuracy of different features on six data sets

seen from Fig. 2f that for LIHC dataset, when the number of features is about 100, the accuracy of the classifier is relatively high. Therefore, the threshold μ_1 corresponding to the LIHC dataset is determined to be 1.53, and the number of selected features is 99.

To sum up, when the NDFS algorithm uses Fisher score to pre-select features in the first stage, the corresponding thresholds and the number of pre-selected features in different datasets are shown in the following Table 4.

For the second stage, to remove redundant features based on similarity measurement and Pareto dominance theory, the

Table 4 Fisher thresholds and the number of pre-selected features in different data sets

Data set	μ_1	Number of pre-selected features
Colon	0.06	414
Leukemia	0.2	406
11_Tumors	1.13	302
Lung	1.17	299
HeadNeck	0.05	287
LIHC	1.53	99

Table 5 Similarity threshold and feature number of optimal feature subset for each dataset

Data set	μ_2	Feature number of optimal feature subset
Colon	0.68	8
Leukemia	0.68	15
11_Tumors	0.68	11
Lung	0.68	4
HeadNeck	0.68	109
LIHC	0.68	3

similarity threshold μ_2 should be determined. In this experiment, cosine similarity was used to measure the similarity between features. Therefore, the closer the value is to 1, the stronger the redundancy between features, and features with high-class correlation are more dominant to other features. Once the similarity threshold has been determined, the Pareto dominance feature corresponding to the feature can be removed. The larger the threshold, the less the Pareto dominance of the feature. The smaller the threshold, the more the Pareto dominating features and the more the deleted features.

Different candidate feature subsets are selected according to threshold μ_1 in Table 4. Calculate the classification accuracy corresponding to different similarity thresholds under candidate feature subsets. After many experiments, the threshold $\mu_2 = 0.68$ is finally determined. At this point, the number of optimal subset features of each dataset is displayed in Table 5.

Therefore, the final number of features selected by this method was determined as follows: Colon-8, Leukemia-15, 11_Tumors-11, Lung-4, HeadNeck-109, and LIHC-3.

Experimental result analysis

In this section, the proposed algorithms are compared with the other six algorithms (Fisher score, MIC, FCBF-MIC, CFR, KNFI, and mRMR) on the six data sets of HeadNeck, Colon, and Leukemia. Different classifiers were used to construct prediction models and evaluate the performance of the feature subsets selected by NDFS.

Table 6 lists the number of features selected by the proposed method and the six comparison algorithms.. According to the experimental setup in Section "Experimental setup", Fisher score, MIC, CFR and mRMR all obtained the same number of dimensional features as NDFS. NDFS first retains only features that are strongly related to the category and more discriminative, greatly reducing the feature dimension, and then performs surprise search based on non-dominated features to retain the best number of features. The second stage of FCBF-MIC uses the MIC between features and features and between features and classes to calculate the approximate Markov blanket. The number of deleted features meeting this condition is less than the number of redundant features abandoned by NDFS using non-dominated feature-guided search. Therefore, in all data sets, the final number of features selected by FCBF-MIC is much higher than that of NDFS.

Tables 7, 8, and 9 show the classification accuracy, F1_score and AUC values of 6 datasets under 7 algorithms and 5 classifiers. Figure 3 shows the ROC curves of the proposed method and the other 6 methods under the SVM classifier. In the ROC curve, the ROC curve of the feature subset obtained by NDFS is closer to the upper left corner, and its performance is better. Compared with other six comparison algorithms, it can be easily explained why the proposed method performs better.

Fisher score, MIC only computes the relationship between features and categories, thus ignoring the relationship between features. NDFS can calculate between features based on removing irrelevant features, eliminating data redundancy, and completing the multi-objective calculation in feature selection. Therefore, compared with Fisher score and MIC algorithm, NDFS algorithm can remove redundant features. As far as the experimental results are concerned, the evaluation indexes of NDFS under the four classifiers in all data sets are superior to Fisher score and mic algorithms. The ACC of RF classifier is slightly lower than that of MIC in three data sets, but the generalization ability of NDFS in different learners is still better than these two algorithms on the whole.

Compared with FCBF-MIC, the classification results of NDFS are better than those of FCBF-MIC on Leukemia, HeadNeck, Colon and Lung datasets. The two algorithms have advantages and disadvantages in using different classifiers on 11_Tumors and LIHC data sets. FCBF-MIC uses SU for correlation analysis in the first stage and MIC for redundancy deletion in the second stage. There is no correlation between the two steps. NDFS can select non-dominated features through Fisher score value, and use two-way search strategy to make full use of the feature information of each calculation.

Compared with CFR, the classification effect of RF is better than that of NDFS only under the Leukemia dataset. However, under the other datasets, the NDFS has an obvious advantage. The mutual information and conditional mutual information used by CFR to calculate the relationship between features is too redundant and has low performance. CFR later used greedy searching strategy, but its final selected number threshold of features was not supported by sufficient solutions. NDFS calculates feature similarity in non-dominant features, and NDFS selects the final optimal subset based on the established similarity threshold.

When selecting features, KNFI focuses more on the influence of the current selected features on the performance

Table 6 The number of features selected by 7 feature selection methods for 6 high-dimensional datasets

Data set	Fisher score	MIC	FCBF-MIC	CFR	KNFI	mRMR	NDFS
Colon	8	8	16	8	3	8	8
Leukemia	15	15	128	15	2	15	15
11_Tumors	11	11	201	11	11	11	11
Lung	4	4	9	4	8	4	4
HeadNeck	109	109	31	109	6	109	109
LIHC	3	3	114	3	2	3	3

Table 7 Classification accuracy, F1 and AUC of Colon and Leukemia datasets under 7 algorithms and different classifiers

Estimator	Method	Colon			Leukemia		
		Accuracy	F1_score	AUC	Accuracy	F1_score	AUC
DT	NDFS	**0.8526**	**0.8972**	**0.8158**	**0.9571**	**0.9647**	**0.9537**
	Fisher score	0.7256	0.758	0.7344	0.8762	0.9026	0.8667
	MIC	0.7731	0.8252	0.735	0.8476	0.8778	0.8445
	FCBF-MIC	0.7372	0.7694	0.7359	0.8067	0.8414	0.8117
	CFR	0.7244	0.7899	0.67	0.8324	0.8661	0.832
	KNFI	0.7923	0.83	0.7891	0.9171	0.937	0.8957
	mRMR	0.7564	0.8025	0.7295	0.8905	0.9131	0.8857
RF	NDFS	0.8833	0.9188	**0.9455**	0.9714	0.9778	0.9819
	Fisher score	0.8692	0.901	0.8356	0.9448	0.9577	0.9826
	MIC	**0.9013**	**0.9231**	0.9275	0.9714	0.9778	**0.9907**
	FCBF-MIC	0.9	0.9193	0.9157	0.9714	0.9778	**0.9907**
	CFR	0.8026	0.8505	0.8732	**0.9724**	**0.9787**	0.9792
	KNFI	0.8538	0.8944	0.9028	0.9295	0.9447	0.9775
	mRMR	0.8526	0.8877	0.9371	0.9581	0.9673	0.9826
SVM	NDFS	**0.9179**	**0.9365**	**0.9768**	**0.9724**	**0.9789**	**0.9949**
	Fisher score	0.8538	0.8892	0.8391	0.9581	0.9673	0.966
	MIC	0.8526	0.8907	0.8904	0.9581	0.9673	0.9788
	FCBF-MIC	0.8526	0.8853	0.9215	**0.9724**	**0.9789**	**0.9949**
	CFR	0.8038	0.8297	0.8293	**0.9724**	0.9777	0.9911
	KNFI	0.8385	0.8816	0.877	0.9438	0.9552	0.973
	mRMR	0.8846	0.9115	0.9336	0.959	0.9694	0.9828
LR	NDFS	**0.9026**	**0.9181**	0.9515	**1**	**1**	**0.9949**
	Fisher score	0.8705	0.9042	0.8773	0.9448	0.9589	0.9704
	MIC	0.8692	0.8981	0.9331	0.9581	0.9673	0.9828
	FCBF-MIC	0.9013	0.9178	0.9462	0.9724	0.9789	0.9907
	CFR	0.7705	0.8106	0.8538	0.9448	0.9542	0.9873
	KNFI	0.8551	0.892	0.8927	0.9305	0.9465	0.9822
	mRMR	0.8859	0.9099	**0.9641**	0.959	0.9694	0.9871
MLP	NDFS	**0.9167**	**0.9349**	**0.9578**	**1**	**1**	**0.9949**
	Fisher score	0.8538	0.8922	0.8806	0.9448	0.9589	0.9745
	MIC	0.8692	0.8981	0.9331	0.9581	0.9673	0.9828
	FCBF-MIC	0.8692	0.8927	0.9273	0.9714	0.9778	0.9907
	CFR	0.8179	0.848	0.8553	0.959	0.9682	0.9911
	KNFI	0.8385	0.8075	0.8422	0.9305	0.9465	0.9822
	mRMR	0.8705	0.8947	0.953	0.9724	0.9789	0.9911

Values are presented as the classification performance index value of the algorithm under the classifier, where the best result on each classifier is shown in bold

Table 8 Classification accuracy, F1 and AUC of 11_Tumors and Lung datasets under 7 algorithms and different classifiers

Estimator	Method	11_Tumors			Lung		
		Accuracy	F1_score	AUC	Accuracy	F1_score	AUC
DT	NDFS	**0.7787**	**0.6896**	**0.9178**	**0.8603**	0.7398	**0.9373**
	Fisher score	0.4095	0.3007	0.7106	0.749	0.5301	0.8385
	MIC	0.7161	0.6055	0.7749	0.7787	0.5018	0.9006
	FCBF-MIC	0.7197	0.6114	0.8964	0.8305	0.6993	0.8995
	CFR	0.6751	0.5558	0.833	0.8223	0.6057	0.8818
	KNFI	0.7329	0.6348	0.8767	0.8135	0.6635	0.8921
	mRMR	0.6312	0.5497	0.7418	0.8462	**0.7540**	0.9060
RF	NDFS	0.8356	0.7312	0.9896	**0.8999**	0.8022	**0.9867**
	Fisher score	0.5141	0.3208	0.8012	0.7935	0.5274	0.9481
	MIC	0.8446	0.7988	0.974	0.8183	0.5387	0.9564
	FCBF-MIC	**0.9051**	**0.8533**	**0.9927**	0.8996	**0.8153**	0.9772
	CFR	0.7412	0.5769	0.9622	0.8726	0.6506	0.9755
	KNFI	0.8421	0.7845	0.9715	0.8823	0.6988	0.9832
	mRMR	0.7285	0.6309	0.6311	0.8712	0.7579	0.9788
SVM	NDFS	0.8727	0.805	**0.9942**	**0.9296**	**0.8744**	0.9784
	Fisher score	0.2952	0.1174	0.7778	0.7789	0.3918	0.9415
	MIC	0.7421	0.5917	0.9726	0.7748	0.3660	0.9634
	FCBF-MIC	**0.9108**	**0.8315**	0.9937	0.9066	0.7842	0.9774
	CFR	0.7228	0.545	0.9372	0.8915	0.6705	0.969
	KNFI	0.7585	0.6296	0.9707	0.8429	0.5909	**0.9828**
	mRMR	0.6148	0.5594	0.5194	0.8958	0.7940	0.9764
LR	NDFS	0.9007	**0.8702**	**0.9939**	**0.9259**	**0.8908**	**0.9803**
	Fisher score	0.3906	0.1824	0.7947	0.7839	0.3727	0.9344
	MIC	0.7806	0.671	0.9708	0.8042	0.4194	0.9709
	FCBF-MIC	**0.9237**	0.8623	0.9937	0.9105	0.7614	0.9789
	CFR	0.7434	0.5712	0.9604	0.8629	0.6092	0.9731
	KNFI	0.7455	0.6414	0.9753	0.8577	0.6842	0.9735
	mRMR	0.7302	0.6111	0.9484	0.8963	0.8243	0.9761
MLP	NDFS	**0.9189**	**0.8893**	**0.9949**	**0.9206**	**0.8963**	**0.982**
	Fisher score	0.4076	0.1918	0.7903	0.7839	0.4124	0.9461
	MIC	0.7888	0.6973	0.9762	0.8328	0.5452	0.9608
	FCBF-MIC	0.9186	0.8653	0.9831	0.8951	0.7929	0.9793
	CFR	0.7443	0.5912	0.959	0.8915	0.6683	0.9737
	KNFI	0.8045	0.7146	0.9781	0.8871	0.7603	0.9818
	mRMR	0.7226	0.6375	0.9624	0.9008	0.8467	0.9766

Values are presented as the classification performance index value of the algorithm under the classifier, where the best result on each classifier is shown in bold

accuracy of the classifier, ignoring the information of the features themselves. In the whole process of NDFS, not only the information of the feature itself is fully used, but also the classification information can be used as a reference for feature correlation threshold selection. The experimental results show that the classification effect of KNFI using DT is better than that of NDFS only under HeadNeck dataset, and NDFS can obtain a feature subset with better classification effect under other dataset classifiers.mRMR uses SVM on the HeadNeck dataset to achieve the best classification indicators, and some indicators on DT, LR, and MLP achieve the best results. F1 is also optimal under the DT classifier on the Lung dataset. But it lacks a competitive advantage under the rest of the datasets. mRMR uses mutual information to calculate feature information, which makes the feature far away and is still highly correlated with classification

Table 9 Classification accuracy, F1 and AUC of HeadNeck and LIHC datasets under 7 algorithms and different classifiers

Estimator	Method	HeadNeck			LIHC		
		Accuracy	F1_score	AUC	Accuracy	F1_score	AUC
DT	NDFS	0.6476	0.6914	0.6418	0.9647	0.8561	0.9261
	Fisher score	0.6095	0.6471	0.6045	0.9788	0.9138	0.9505
	MIC	0.6571	0.6719	0.6554	0.9788	0.9136	**0.9599**
	FCBF-MIC	0.6762	0.6791	0.6773	**0.9811**	**0.9148**	0.9432
	CFR	0.5905	0.6336	0.5864	0.9246	0.6514	0.8005
	KNFI	**0.6857**	0.7008	**0.6848**	0.9694	0.8699	0.9289
	mRMR	0.6762	**0.702**	0.6709	0.9717	0.8823	0.93
RF	NDFS	0.8381	0.8466	**0.9138**	**0.9859**	**0.9433**	0.9933
	Fisher score	0.781	0.8025	0.864	0.9741	0.8987	0.9818
	MIC	**0.8476**	**0.8612**	0.9033	**0.9859**	0.9407	**0.9941**
	FCBF-MIC	0.8	0.8263	0.8928	0.9835	0.9331	**0.9941**
	CFR	0.7714	0.7908	0.8912	0.9292	0.6284	0.9614
	KNFI	0.7714	0.7816	0.8466	0.9811	0.9242	0.9919
	mRMR	0.8286	0.846	0.9125	0.9787	0.9083	0.9937
SVM	NDFS	**0.9143**	0.9146	0.9595	**0.9835**	**0.9361**	0.9915
	Fisher score	0.8476	0.8519	0.934	0.9788	0.9169	0.9719
	MIC	0.7429	0.7603	0.8398	0.9741	0.9028	**0.9937**
	FCBF-MIC	0.8571	0.8654	0.9285	**0.9835**	0.9357	0.9931
	CFR	0.7714	0.768	0.8416	0.9552	0.8320	0.9663
	KNFI	0.781	0.806	0.821	0.9764	0.9090	0.9853
	mRMR	**0.9143**	**0.9157**	**0.9624**	0.9717	0.8921	0.9909
LR	NDFS	**0.9143**	**0.9146**	0.9604	0.9788	0.9157	**0.9945**
	Fisher score	0.8952	0.9031	0.9362	0.9788	0.9195	0.9933
	MIC	0.7429	0.7745	0.8581	0.9765	0.9133	0.9941
	FCBF-MIC	0.8286	0.8327	0.9156	0.9788	0.919	0.9929
	CFR	0.6667	0.687	0.7736	0.9434	0.7668	0.9822
	KNFI	0.7143	0.7241	0.7922	0.9764	0.9074	0.9913
	mRMR	0.9048	0.9126	**0.9672**	**0.9811**	**0.9254**	0.9935
MLP	NDFS	**0.9238**	**0.9251**	0.9656	**0.9859**	**0.9466**	**0.9943**
	Fisher score	0.8762	0.876	0.9316	0.9788	0.9195	0.9931
	MIC	0.7619	0.7789	0.8243	0.9788	0.9206	0.9937
	FCBF-MIC	0.8381	0.8478	0.9152	0.9787	0.9176	0.9939
	CFR	0.7429	0.76	0.7861	0.9505	0.8114	0.9818
	KNFI	0.819	0.8373	0.8074	0.9764	0.9100	0.9929
	mRMR	0.9143	0.9181	**0.9702**	0.9811	0.9231	0.9927

Values are presented as the classification performance index value of the algorithm under the classifier, where the best result on each classifier is shown in bold

variables. NDFS calculates the feature similarity by considering the relationship between features to remove the disposable features, to generate a reliable feature subset.

Overall, the proposed algorithm uses NDFS in six datasets to select effective features for current problems. It has better performance in both binary and multi-classification tasks, and has good generalization ability on different classifiers.

To further compare the differences among the algorithms, Friedman nonparametric test was used to calculate, and the average ranking of the algorithms was used to judge whether there were significant differences between the algorithms. The average ranking of 7 algorithms in 6 datasets was calculated and compared with the f-distribution critical value with a confidence degree of 0.1 to obtain Table 10. It can be

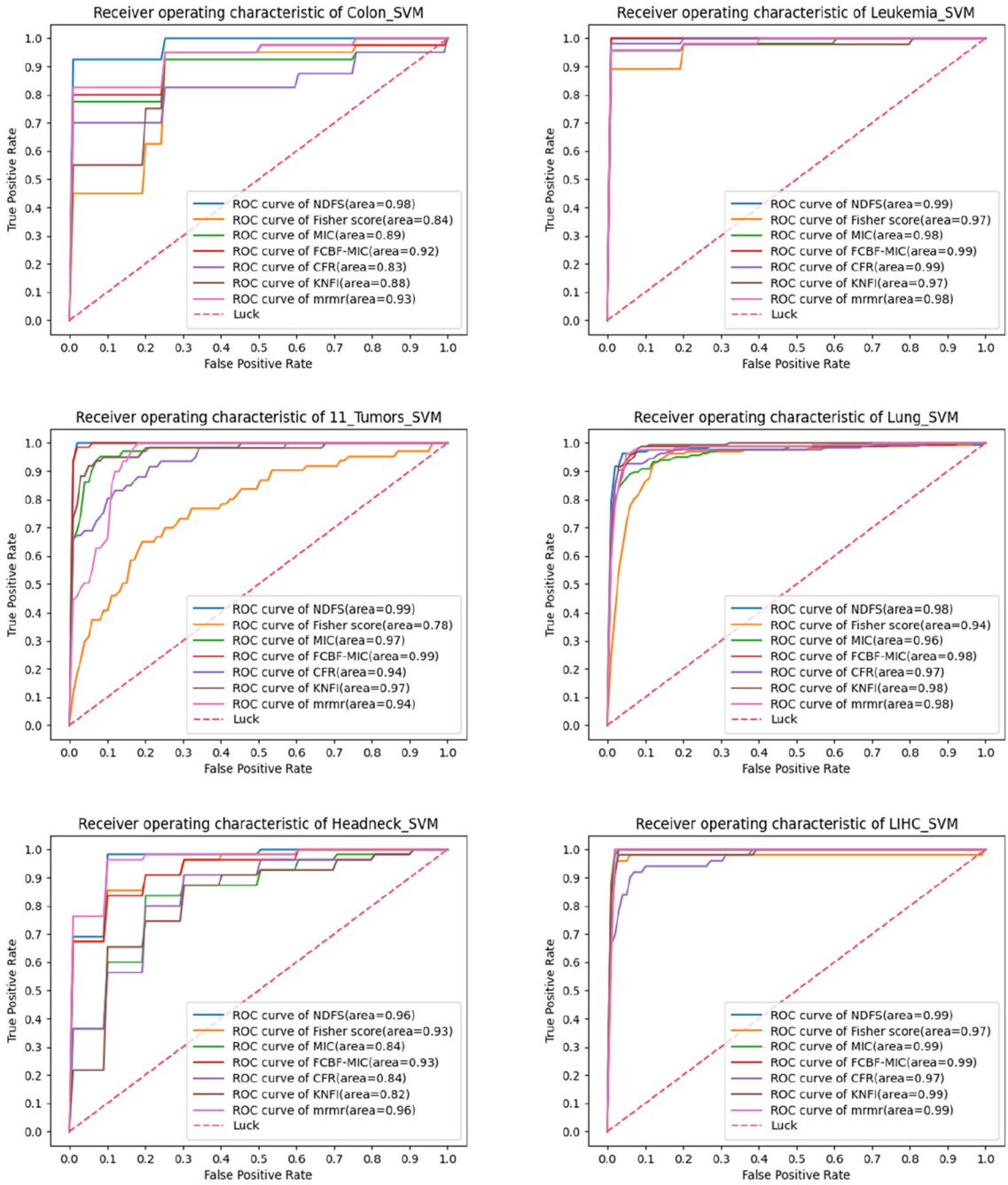

Fig. 3 Receiver Operating Characteristic of six feature selection methods on Colon and Leukemia datasets under SVM classifier

seen from Table 10 show that for five classifiers, T_F of classification accuracy and T_F of F1_score are both greater than the corresponding critical value, indicating significant differences between algorithms. Therefore, the proposed NDFS algorithm has good performance.

NDFS algorithm, in the best case, to be selected features are the key feature subset of the first feature dominating features. The time complexity is $O(n)$. In the worst case, the unselected features are non-dominated features of the key

Table 10 Friedman test

Estimator	T_F		Critical value
	Accuracy	F1-score	
DT	2.7410	3.9750	1.980
RF	3.8091	5.0191	
SVM	5.0877	5.5123	
LR	8.0921	8.4172	
MLP	7.2170	4.5226	

Table 11 Average running time(s) of 7 algorithms from 6 datasets

Data set	Fisher score	MIC	FCBF-MIC	CFR	KNFI	mRMR	NDFS
Colon	0.493	0.456	0.825	101.863	504.346	8.350	0.662
Leukemia	1.755	1.888	2.517	831.401	1742.315	18.812	1.907
11_Tumors	52.417	61.836	101.712	1417.058	3197.736	26.385	53.711
Lung	39.901	31.937	39.677	19197.781	3365.183	14.868	34.975
HeadNeck	7.760	7.644	10.043	13914.121	3120.462	160.906	8.719
LIHC	95.957	83.268	161.907	1084.942	13730.563	86.745	119.15

feature subset, and the time complexity is $O(n^2)$. Since the NDFS algorithm can delete many irrelevant features first, it will reduce the Pareto dominance feature calculation scale and improve the overall speed of the method. Therefore, NDFS has relatively fast operating efficiency. The feature selection experiments on each data set were repeated 10 times, and then the mean was calculated as the estimation of the running time of feature selection. The final experimental results are shown in Table 11.

NDFS algorithm is a pre-selection based on Fisher score, which is higher than Fisher score and MIC algorithm in running time. However, the classification experiments show that the classification performance is poor, and the two algorithms eventually contain high redundancy features, so the proposed algorithm time is relatively acceptable. It can be seen from the table that compared with FCBF-MIC, CFR, and KNFI algorithms, NDFS has a lower running time on six datasets. Compared with mRMR algorithm, NDFS has lower running time on three datasets.

In summary, when the same number of feature subsets is selected, the classification performance of feature subsets selected by NDFS algorithm is better than those of Fisher score, MIC, CFR, and mRMR under most classifiers. NDFS selects important features and eliminates redundant features, which significantly preserves useful feature information. Compared with the FCBF-MIC and KNFI algorithm, NDFS adopts the distance-based measure more accurately than the probability value of information theory, so the performance of the selected feature subset is better.

Conclusion

In this paper, a feature selection using non-dominant features-guided search is proposed. Fisher score algorithm is used to measure the category correlation of features. Cosine similarity based on geometric distance standard is used to measure the similarity between features. Specifically, the algorithm combines the Pareto dominance theory to gradually remove the Pareto dominance feature (redundant feature) of the largest category correlation feature. A feature subset with maximum correlation and minimum redundancy is obtained.

This algorithm uses the fast and effective characteristics of the Fisher score algorithm to select related features. It makes up for the deficiency of Fisher score that does not consider the correlation between features. The proposed method is compared to six competing feature selection methods on six real-world data sets. This approach has better classification performance than Fisher score, MIC, CFR, and mRMR algorithms and does not only consider the category correlation or feature redundancy of features. Compared with the FCBF-MIC and KNFI algorithms, it can obtain the feature subset with better classification ability while accelerating the algorithm execution efficiency. In light of the above experimental results show that NDFS method outperforms other compared feature selection methods.

NDFS has been able to extract features with low redundancy and affecting gene category. It is also worth studying to analyze the genes that affect the disease from the selected genes. Further work will establish an interpretable association model of selected features and final results to provide a scientific basis for personalized diagnosis and treatment.

Funding National Key R&D Program of China (Item NO: 2019YFC0121502), Project supported by the Special Fund of Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Ram PK, Kuila P (2019) Feature selection from microarray data: genetic algorithm based approach[J]. J Inform Optim Sci 40(8):1599–1610
2. Lim K, Li Z, Choi KP, Wong L (2015) A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small [J]. J Bioinform Computational Biol 13(4):1550018–1550018
3. Xue Y, Xue B, Zhang M (2019) Self-adaptive particle swarm optimization for large-scale feature selection in classification[J]. ACM Trans Knowl Discov from Data (TKDD) 13(5):1–27
4. Hambali MA, Oladele TO, Adewole KS (2020) Microarray cancer feature selection: review, challenges and research directions[J]. Int J Cogn Computing Eng 1(1):78–97
5. Rui Z, Feiping N et al (2019) Feature selection with multi-view data: a survey ScienceDirect[J]. Int J Inform Fusion 50:158–167
6. Manikandan G, Susi E, Abirami S (2019) Flexible-fuzzy mutual information based feature selection on high dimensional data[C]//2018 Tenth International Conference on Advanced Computing (ICoAC). IEEE
7. Nakariyakul S (2018) High-dimensional hybrid feature selection using interaction information-guided search[J]. Knowl-Based Syst 145(1):59–66
8. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
9. Yu L, Liu H (2004) Eficient feature selection via analysis of relevance and redundancy[J]. J Mach Learn Res 5(12):1205–1224
10. Gao T, Ji Q (2016) Efficient Markov blanket discovery and its application[J]. IEEE Trans Cybern 47(5):1169–1179
11. Jo I, Lee S, Sejong Oh (2019) Improved measures of redundancy and relevance for mRMR feature selection[J]. Computers 8(2):42–42
12. Li S et al (2020) Feature selection for high dimensional data using weighted K-nearest neighbors and genetic algorithm. IEEE Access 8(8):139512–139528
13. Cai J, Luo J, Wang S et al (2018) Feature selection in machine learning: a new perspective[J]. Neurocomputing 300(26):70–79
14. Kira K, Rendell LA (1992) A practical approach to feature selection[J]. Proceedings of the Ninth International Workshop on Machine Learning (ML 1992), Aberdeen, Scotland, UK, July 1–3
15. Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF[J]. Mach Learn, 53(1–2)
16. Gu Q, Li Z, Han J (2012) Generalized fisher score for feature selection[J]
17. Reshef DN, Reshef YA, Finucane HK et al (2011) Detecting novel associations in large data sets[J]. Science 334(6062):1518–1524
18. Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning. 359−366
19. Gao W, Liang Hu, Zhang P, He J (2018) Feature selection considering the composition of feature relevancy[J]. Pattern Recogn Lett 112:70–74
20. Estevez PA, Tesmer M, Perez CA et al (2009) Normalized mutual information feature selection[J]. IEEE Trans Neural Netw 20(2):189–201
21. Javed K, Babri HA, Saeed M (2012) Feature selection based on class-dependent densities for high-dimensional binary data[J]. IEEE Trans Knowl Data Eng 24(3):465–477
22. Sun G, Song Z, Liu J et al (2017) Feature selection method based on maximum information coefficient and approximate markov blanket[J]. Acta Automatica Sinica (in Chinese) 43(05):795–805
23. Zhang L, Wang C et al (2018) A feature selection algorithm for maximum relevance minimum redundancy using approximate markov blanket[J]. J Xi'an Jiaotong Univ (in Chinese) 52(10):141–145
24. Thejas GS, Joshi SR, Iyengar SS et al (2019) Mini-batch normalized mutual information: a hybrid feature selection method[J]. IEEE Access 7(99):116875–116885
25. Xue B, Zhang MJ, Browne WN (2013) Particle swarm optimization for feature selection in classification: a multi-objective approach[J]. IEEE Trans Cybern 43(6):1656–1671
26. Zhou Yu, Kang J, Guo H (2020) Many-objective optimization of feature selection based on two-level particle cooperation[J]. Inf Sci 532(532):91–109
27. Saha S, Ghosh M, Ghosh S, Sen S, Singh PK, Geem ZW, Sarkar R (2020) Feature selection for facial emotion recognition using cosine similarity-based harmony search algorithm[J]. Appl Sci 10(8):2816
28. Walter V, Yin X, Wilkerson MD et al (2013) Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes[J]. PLoS ONE 8(2):e56823