MATHEMATICAL MODELING AND COMPUTING, Vol. 10, No. 3, pp. 693–702 (2023)

PROMETHEE filter-based method for microarray gene expression data

Ouaderhman T., Aaboub F., Chamlal H.

Department of Mathematics and Computer Science, Fundamental and Applied Mathematics Laboratory, Faculty of Sciences Ain Chock, Hassan II University, Casablanca, Morocco

(Received 17 February 2023; Accepted 6 July 2023)

Gene expression datasets have been successfully applied for a variety of purposes, including cancer classification. The challenges faced in developing effective classifiers for expression datasets are high dimensionality and over-fitting. Gene selection is an effective and efficient method to overcome these challenges and improve the predictive accuracy of a classifier. Based on PROMETHEE, this paper introduces a multi-filter ensemble approach by integrating the results of two potential filters namely MaCΨ-filter and PCRWG-filter to pre-select the most informative genes. Experiments were conducted on nine microarray datasets to demonstrate the performance of the proposed method.

Keywords: *filter; classification; selection; microarray data; PROMETHEE.*

2010 MSC: 62-07, 97K80, 68T20, 65C60 **DOI:** 10.23939/mmc2023.03.693

1. Introduction

The dimensionality curse affects certain dataset types, including microarray gene expression datasets. Gene or feature selection is thus a crucial preprocessing stage required before analyzing a microarray dataset. Through the selection of the most important genes and the elimination of all the less significant ones, this phase aids in the reduction of the gene space.

Many feature selection methods are presented in the scientific literature. Three main categories of feature selection methods are filter, wrapper, and hybrid methods [1]. On the one hand, filter-based techniques evaluate the features of a dataset solely based on their own merits, unrelated to a classifier. Each feature is then given a score, and the top N ranked features are chosen as the most significant ones while the remaining features are disregarded. The filter-based algorithms are further divided into two types: univariate and multivariate strategies, depending on whether or not the interaction among features is taken into account. Each feature is evaluated separately by univariate techniques, while numerous features are evaluated concurrently by multivariate techniques. In contrast to filter-based processes, the wrapper methods need a learning algorithm to evaluate the classification effectiveness of the selected features. The hybrid methodologies, on the other hand, attempt to combine a filter method with a wrapper method in order to benefit from the advantages of both. Numerous feature selection techniques based on filters have been suggested by the researchers. Some of the well-known univariate filter-based techniques include Mutual Information [2], Information Gain (IG) [3], Gain Ratio (GR) [4], Fisher score [5], Laplacien score [6], Chi-square statistic [7], and Symmetric Uncertainty (SU) [8]. Moreover, some of the common multivariate filter-based techniques include Fast Correlation-Based Filter [9], ReliefF [10], and minimal Redundancy Maximal Relevance [11].

An innovative multi-filter ensemble method for gene selection is presented in this paper. The recommended approach combines two phases. The dimensionality of genes is decreased in the first stage to form the candidate subset of genes by only selecting the genes that are the most d relevant. The final rank produced by the proposed method is derived from two rankings determined by the MaCΨ-filter and PCRWG-filter using the PROMETHEE method. Using the Sequential Forward Selection (SFS) technique on the candidate subset, the final subset of genes and its classification accuracy are obtained in the final stage.

2. Preliminary concepts

This section presents some concepts that will be used to define the proposed multi-filter ensemble approach for feature selection, including the relevancy, redundancy, and complementarity measures, as well as the Preference Ranking Organization METHod for Enrichment Evaluations (PROMETHEE).

Let P_{X_i}, P_{X_j}, and P_Y represent three preordonances that are induced by two description variables X_i, X_j, and a class variable Y, respectively. These preordonances are designated, respectively, by the ternary codings $T_{P_{X_i}}$, $T_{P_{X_j}}$, and T_{P_Y}.

Relevance measure [12]. The relevance measure, denoted by Ψ, of a given variable X_i with respect to the class variable Y can be calculated using the ψ_{cor} metric using the following equation:

$$\begin{aligned}\Psi(X_i) &= \psi_{cor}(P_{X_i}, P_Y)\\ &= \operatorname{cor}(T_{PX_i}, T_{P_Y})\\ &= \frac{\operatorname{cov}(T_{P_{X_i}}, T_{P_Y})}{\sqrt{\operatorname{var}(T_{P_{X_i}})}\sqrt{\operatorname{var}(T_{P_Y})}},\end{aligned}$$

where cor, cov, and var indicate the correlation, covariance, and variance measures, respectively.

The more significant the variable X_i, the more relevant it is to the class variable.

Redundancy measure [12]. The most important variables are not only those that are relevant; they should also not be redundant. In this paper, to evaluate the redundancy of a variable X_i to another variable X_j, the following measure can be used:

$$\Psi_{\cdot}(X_i, X_j) = \psi_{cor}(P_{X_i}, P_Y)_{\cdot P_{X_j}}. \qquad (1)$$

Complementarity measure [13, 14]. The following metric can be used to assess the complementarity of a subset of variables X_i and X_j with regard to the class variable Y:

$$\Psi_{\omega}(X_i, X_j) = W(r_i, r_j, r), \qquad (2)$$

where r_i, r_j, and r are the ranks induced by the preordonances P_{X_i}, P_{X_j}, and P_Y, respectively.

PROMETHEE strategy. Preference Ranking Organization METHod for Enrichment Evaluations (PROMETHEE) is a decision making technique [15] that helps in decision-making when there are several viable solutions to a problem. The PROMETHEE is a Multi Attribute Decision Making (MADM) method that can be employed when there are several alternatives available and one should be chosen. Consider a situation where there are M attributes $\{R_1, R_2, R_3, \ldots, R_M\}$ and N alternatives $\{AL_1, AL_2, AL_3, \ldots, AL_N\}$. MADM problems is modeled by a decision matrix, which is defined in Table 1.

Table 1. Decision matrix for PROMETHEE method.

Alternatives	Attributes				
	R_1 (ω_1)	R_1 (ω_2)	R_3 (ω_3)	$\ldots$ $(\ldots)$	R_M (ω_M)
AL_1	e_{11}	e_{12}	e_{13}	$\ldots$	e_{1M}
AL_2	e_{21}	e_{22}	e_{23}	$\ldots$	e_{2M}
AL_3	e_{31}	e_{32}	e_{33}	$\ldots$	e_{3M}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
AL_N	e_{N1}	e_{N2}	e_{N3}	$\ldots$	e_{NM}

3. Multi-filter ensemble approach for gene selection

This section outlines the full procedure of the suggested multi-filter ensemble approach for selecting the most important genes in the microarray datasets. As a pre-processing phase of the suggested methodology, the dimensionality of the microarray datasets should first be reduced, since this type of dataset contains thousands of genes. Consequently, the previously established relevance measure is used in this work to perform the reduction. In order to form a candidate subset of genes, the relevance value of each gene is calculated, and the top d relevant genes are chosen, while the remaining genes are disregarded. It should be noted that X_{i_1} designates the gene with the highest level of relevance. The second stage then employs two filter-based methodologies that are MaCΨ-filter [12] and PCRWG-filter [13] to determine the significance of the genes in the candidate subset. The first filter strategy combines relevance and redundancy metrics to select the relevant and non-redundant genes, while the second combines relevance and complementarity metrics to choose the most significant subset of genes. These two filter methods are defined by two measures that assign two scores to each gene X_i in the following ways:

$$Score\,1(X_i) = \frac{\Psi(X_i)}{2 - \Psi_{.}(X_i, X_{i_1})}, \tag{3}$$

$$Score\,2(X_i) = \frac{\Psi(X_i)}{2 - \Psi_{\omega}(X_i, X_{i_1})}. \tag{4}$$

The first score aims to choose the most relevant gene, which is distinct from the first gene chosen (X_{i_1}). The second score, on the other hand, aims to choose a subset $\{X_{i_1}, X_i\}$ that can provide the most information about the class label. In this framework, two different rankings of the genes of the candidate subset are determined using these two filter-based techniques. The obtained rankings will be aggregated using the PROMETHEE method to identify the final ranking for the d genes. Finally, the sequential forward selection is applied on the obtained ranked subset to identify the final subset of genes with the highest classification accuracy. Figure 1 summarizes the workflow of the recommended multi-filter ensemble approach.

4. Experimental studies

This section provides the experiments that were performed to evaluate the effectiveness of the suggested strategy in enhancing classification performance and escaping the dimensionality curse. Section 4.1 first describes the tested datasets and the computing environment. After that, in Section 4.2, the employed classifiers and the evaluation metrics that rate classification performances are presented. Finally, Section 4.3 depicts the obtained results along with an analysis.

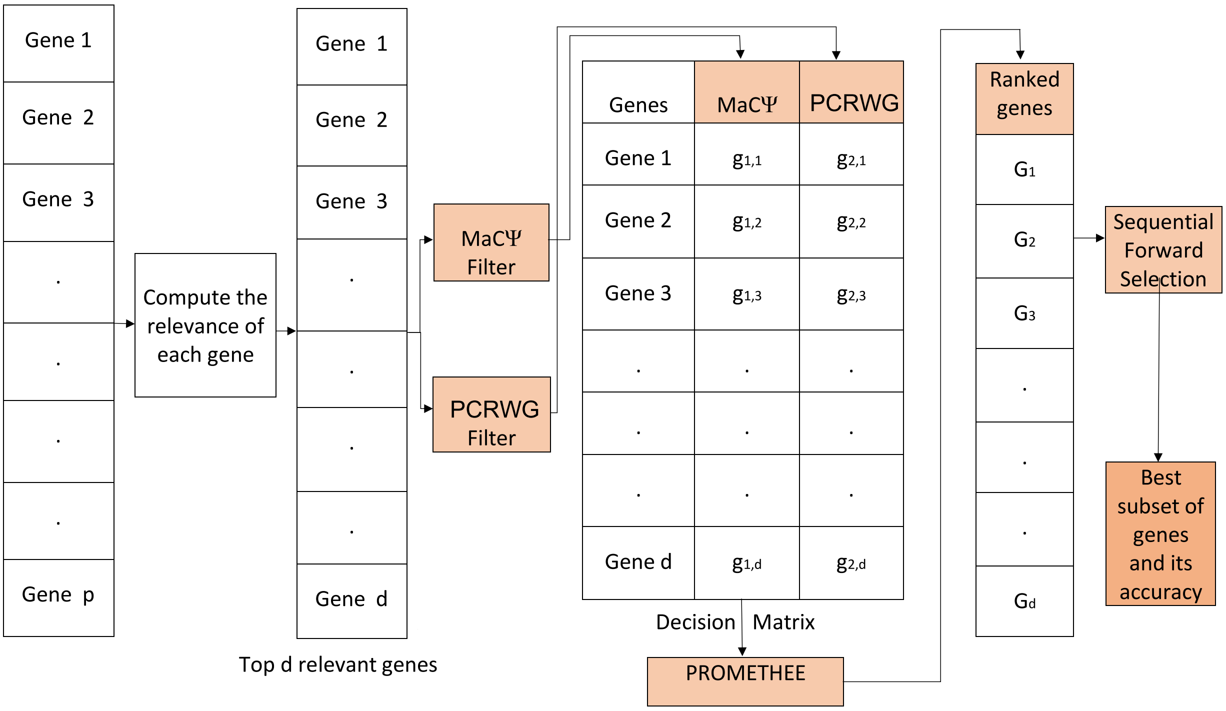

Fig. 1. Workflow of the proposed methodology.

4.1. Dataset and computing environment description

The performance of the proposed approach is examined through a collection of studies on nine widely used microarray gene expression datasets with varying feature sizes, namely Breast, CNS, Colon,

Table 2. Description of the microarray gene expression datasets.

N^o.	Datasets	#Observations	#Features	#Classes
1	Breast	78	4348	2
2	CNS	60	7129	2
3	Colon	62	2000	2
4	Leukemia	72	7128	2
5	Lymphoma	62	4026	3
6	Prostate	102	12600	2
7	SRBCT	83	2308	4
8	Westbc	49	7129	2
9	9_Tumors	60	5726	9

Leukemia, Lymphoma, Prostate, SRBCT, Westbc, and 9_Tumors. Six of the nine datasets have binary classes, and three of them are multi-class. The traits of these datasets, including the number of observations, the number of features, and the number of classes, are summarized in Table 2.

On the other hand, the experimental studies were conducted on a personal computer (Microsoft Windows 10, 11th Gen Intel(R) Core(TM) i7-11850H @ 2.50GHz, 32.0 Gb of RAM, and a 64-bit operating system). Moreover, all of the experiments were carried out using the "R version 4.2.1" computing environment.

4.2. Classifiers and performance metrics

Support Vector Machine (SVM) [16], Naive Bayes (NB) [17], and k-Nearest Neighbors (kNN) [18] are three popular learning algorithms that will be utilized in this paper to analyze the classification performance of the feature selection procedures. Furthermore, classification accuracy, sensitivity, precision, and F-measure are four classification performance metrics that will be employed to rate the picked feature subsets. These evaluation metrics are described in Table 3, where T_P, T_N, F_P, and F_N stand for true positive, true negative, false positive, and false negative, respectively.

Table 3. Classification performance metrics.

Metrics	Formulas	Metrics	Formulas
Accuracy	$\frac{T_P+T_N}{T_P+T_N+F_P+F_N}$	Precision	$\frac{T_N}{T_N+F_P}$
Sensitivity	$\frac{T_P}{T_P+F_N}$	F-measure	$\frac{2\times precision\times sensitivity}{precision+sensitivity}$

Another performance indicator for assessing the effectiveness of feature selection methods is the number of features that are selected.

4.3. Experimental results and discussion

To demonstrate the efficiency of the recommended algorithm in retrieving the most significant features and eliminating the peripheral ones, two experiments are conducted. In terms of classification accuracy and the number of selected features, the first experiment contrasts the performance of the suggested technique with that of four traditional filter-based algorithms. Moreover, the second experiment compares the classification results of the proposed method to those obtained without feature selection (without FS). The value of d in the proposed procedure is kept constant throughout all experiments at 100. Furthermore, all results are estimated from the average of 500 times ten-fold cross-validation for accurate evaluation.

Comparison with some other filter-based feature selection algorithms. The first experiment aims to compare the performance of the proposed technique with that of four well-known filter-based approaches. These four methodologies are mRMR [19], GR [4], IG [3], and Fisher [5]. Figures 2–10 illustrate how the classification accuracy of the five filter-based approaches changes on nine datasets using the SVM, NB, and kNN classifiers as the number of top-ranked features based on the order evaluated by each method increases from 2 to 40.

According to the analysis of Figures 2–10, it can be seen that the general trend of all compared feature selection strategies is analogous, indicating that a number of genes not exceeding 40 is typically required to achieve the highest classification accuracy on the nine datasets using each of the SVM, NB, and kNN classifiers. Furthermore, it is clear that the classification accuracy of the five filter-based approaches does not increase monotonically as the number of genes increases. In other words, the trend of classification accuracy cannot be predicted as the number of selected genes rises.

On the other hand, Tables 4–6 display the highest level of classification accuracy attained by all five processes on nine datasets using the SVM, NB, and kNN classifiers, respectively. Similarly, Tables 7–9

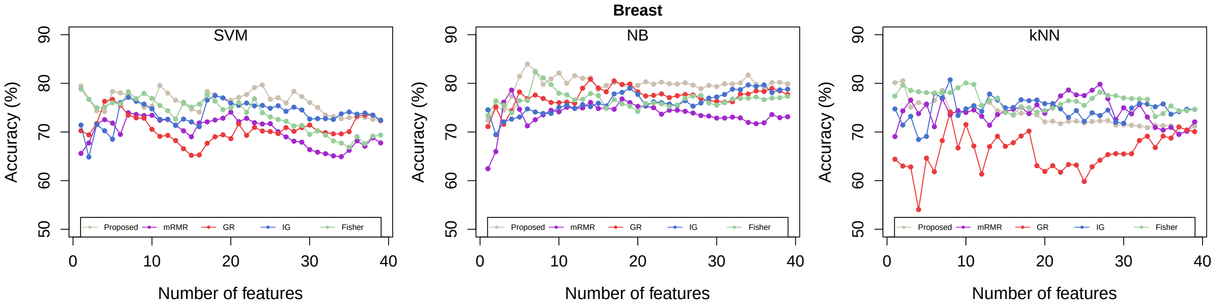

Fig. 2. Classification accuracy versus the number of features for the Breast dataset.

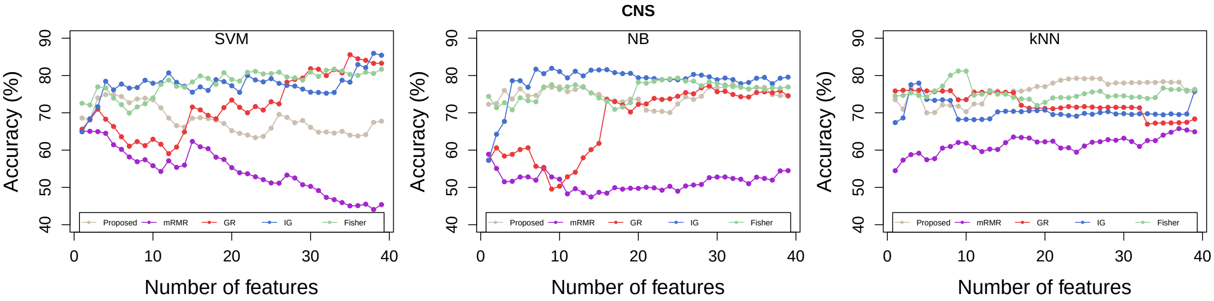

Fig. 3. Classification accuracy versus the number of features for the CNS dataset.

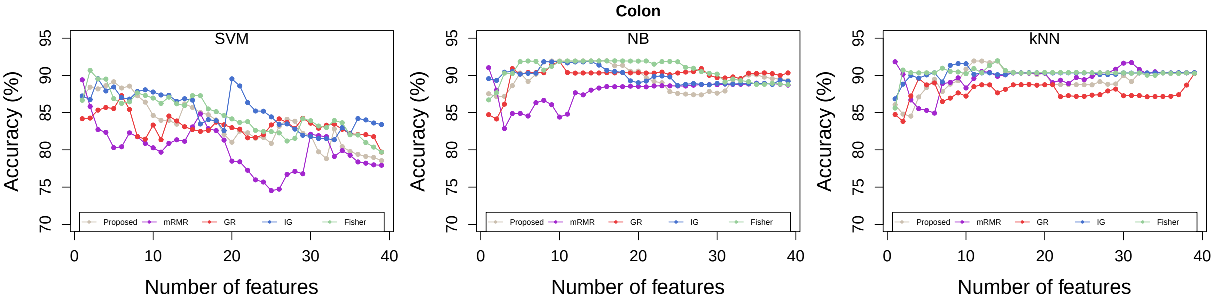

Fig. 4. Classification accuracy versus the number of features for the Colon dataset.

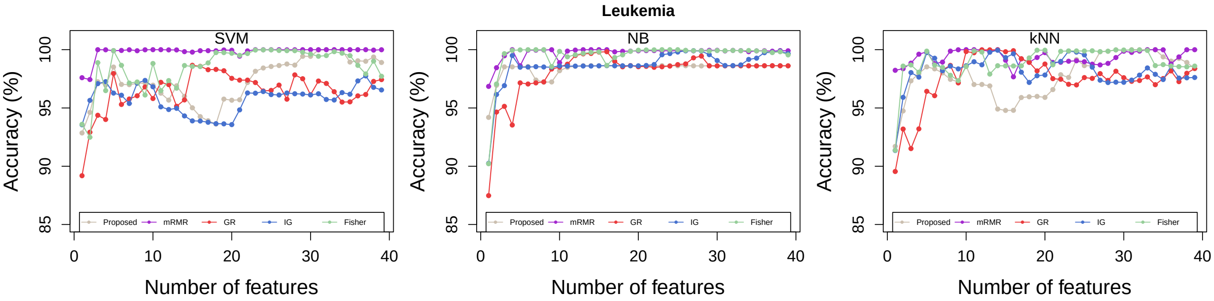

Fig. 5. Classification accuracy versus the number of features for the Leukemia dataset.

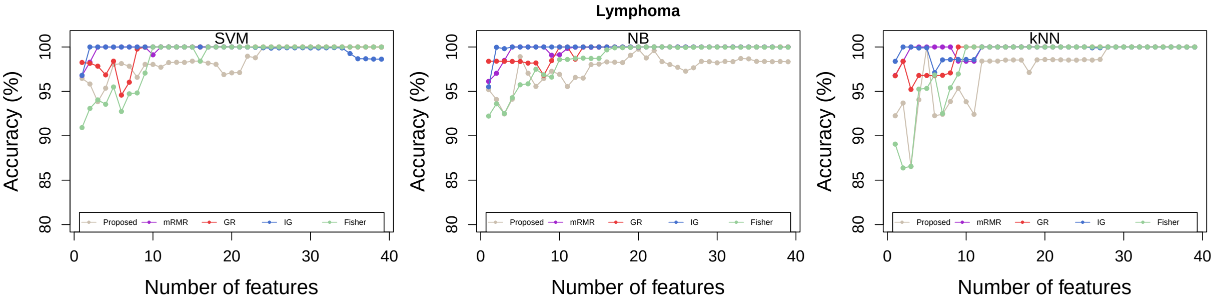

Fig. 6. Classification accuracy versus the number of features for the Lymphoma dataset.

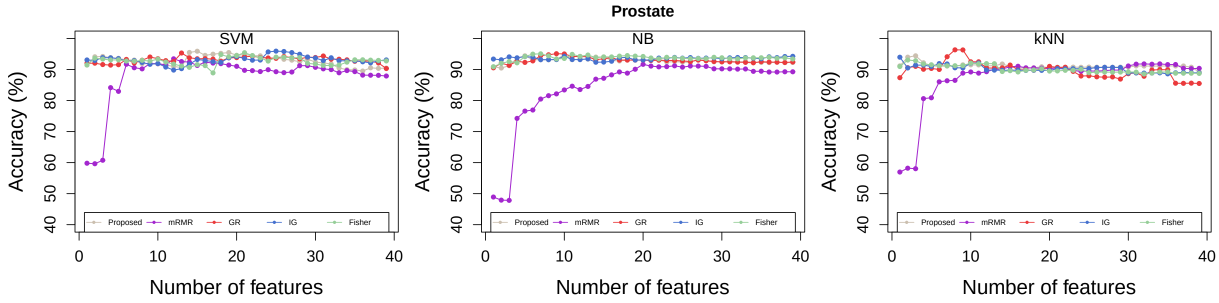

Fig. 7. Classification accuracy versus the number of features for the Prostate dataset.

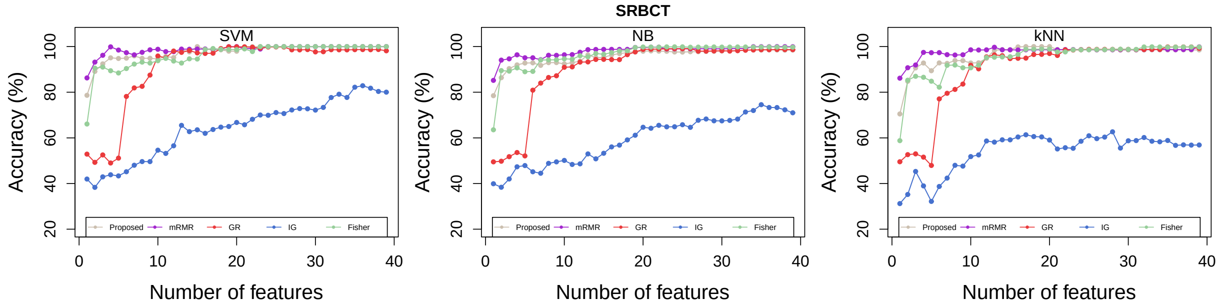

Fig. 8. Classification accuracy versus the number of features for the SRBCT dataset.

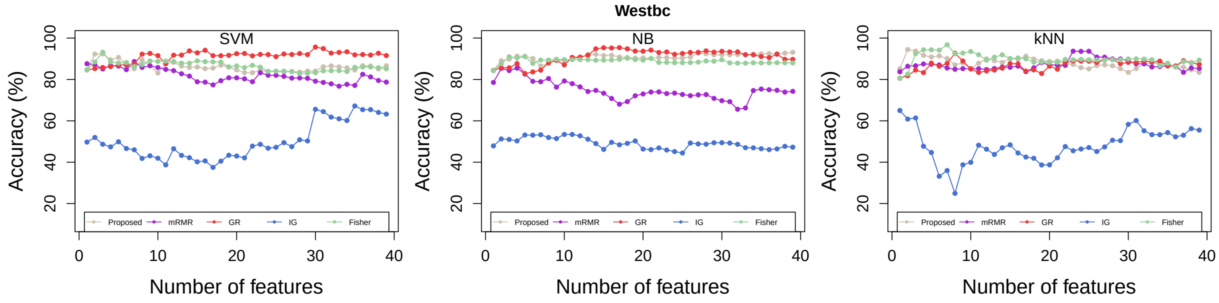

Fig. 9. Classification accuracy versus the number of features for the Westbc dataset.

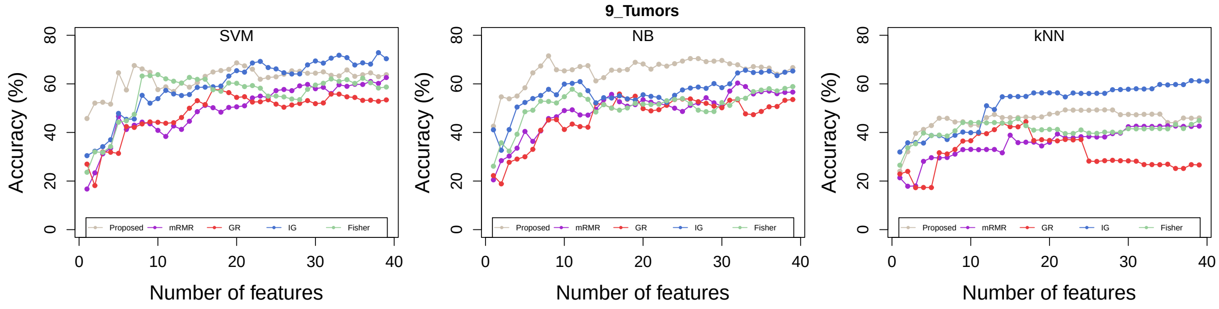

Fig. 10. Classification accuracy versus the number of features for the 9_Tumors dataset.

reveal the corresponding number of selected genes that yield the maximum classification accuracy for each dataset. The final column in the tables displays the average values (the average classification accuracy and the average number of genes) across all datasets.

Table 4. Testing accuracies attained using the SVM classifier by the proposed, mRMR, GR, and IG methods.

Datasets	Breast	CNS	Colon	Leukemia	Lymphoma	Prostate	SRBCT	Westbc	9_Tumors	Average
Proposed	**79.67**	74.86	89.14	**100**	**100**	**95.90**	**100**	92.35	**68.65**	88.95
mRMR	74.08	65.04	89.40	**100**	**100**	93.46	**100**	88.67	62.61	85.92
GR	76.69	85.58	87.28	98.66	**100**	95.32	99.91	**95.65**	57.82	88.55
IG	77.47	**85.95**	89.56	97.75	**100**	95.88	82.85	67.18	72.85	85.50
Fisher	78.83	81.72	**90.67**	99.98	**100**	95.43	**100**	93.15	63.80	**89.29**

Table 5. Testing accuracies attained using the NB classifier by the proposed, mRMR, GR, and IG methods.

Datasets	Breast	CNS	Colon	Leukemia	Lymphoma	Prostate	SRBCT	Westbc	9_Tumors	Average
Proposed	**83.98**	77.63	**91.96**	98.63	99.75	94.77	98.60	93.12	**71.52**	**89.99**
mRMR	78.61	58.89	91.02	**100**	**100**	91.68	**99.98**	85.47	60.34	85.11
GR	80.84	77.23	91.86	99.82	**100**	**95.10**	99.15	**95.37**	55.83	88.36
IG	79.69	**81.93**	91.87	99.91	**100**	94.28	74.55	53.42	65.62	82.35
Fisher	82.29	79.34	91.95	**100**	**100**	95.06	99.88	91.00	58.88	88.71

Table 6. Testing accuracies attained using the kNN classifier by the proposed, mRMR, GR, and IG methods.

Datasets	Breast	CNS	Colon	Leukemia	Lymphoma	Prostate	SRBCT	Westbc	9_Tumors	Average
Proposed	80.59	79.25	**91.93**	**100**	**100**	94.43	**100**	94.46	49.35	**87.78**
mRMR	79.81	65.76	91.82	**100**	**100**	91.82	99.87	93.67	42.74	85.05
GR	74.03	76.07	90.24	99.99	**100**	**96.35**	99.85	92.68	44.42	85.96
IG	**80.62**	77.97	91.59	99.95	**100**	93.97	62.66	64.99	**61.32**	81.45
Fisher	80.08	**82.75**	91.92	99.98	**100**	93.06	99.88	**96.75**	45.56	**87.78**

Table 7. Number of selected genes by the proposed, mRMR, GR, and IG methods using the SVM classifier.

Datasets	Breast	CNS	Colon	Leukemia	Lymphoma	Prostate	SRBCT	Westbc	9_Tumors	Average
Proposed	12	5	6	33	30	16	37	3	21	18.11
mRMR	21	3	2	4	5	13	25	8	40	13.44
GR	6	36	7	16	11	14	20	31	18	17.67
IG	19	39	4	38	3	26	37	36	39	26.78
Fisher	2	34	3	26	11	22	24	4	11	15.22

Table 8. Number of selected genes by the proposed, mRMR, GR, and IG methods using the NB classifier.

Datasets	Breast	CNS	Colon	Leukemia	Lymphoma	Prostate	SRBCT	Westbc	9_Tumors	Average
Proposed	7	33	10	4	21	11	37	40	9	19.11
mRMR	5	2	2	5	5	21	38	5	33	12.89
GR	15	30	11	17	11	10	27	18	18	17.44
IG	37	10	15	28	5	11	36	11	34	20.78
Fisher	8	26	18	9	21	8	25	4	40	17.68

Table 9. Number of selected genes by the proposed, mRMR, GR, and IG methods using the kNN classifier.

Datasets	Breast	CNS	Colon	Leukemia	Lymphoma	Prostate	SRBCT	Westbc	9_Tumors	Average
Proposed	3	25	12	32	29	4	18	3	28	17.11
mRMR	28	38	2	11	4	35	40	24	36	24.22
GR	9	5	40	14	10	9	36	9	18	16.67
IG	9	5	10	15	3	2	29	2	39	12.67
Fisher	11	37	15	33	11	3	34	8	17	18.78

The examination of the experimental results shows that the proposed filter-based method can produce satisfactory classification accuracy on all the tested datasets (on average, $88.95\%, 89.99\%$, and 87.78% when using the SVM, NB, and kNN classifiers, respectively) with only a small number of genes (on average, 18.11, 19.11, and 17.11 genes when using the SVM, NB, and kNN classifiers, respectively). On the one hand, using the SVM classifier, the mRMR, GR, IG, and Fisher algorithms can each achieve the highest classification accuracy on 3, 2, 2, and 3 datasets, respectively. In contrast, the suggested technique reaches the maximum classification accuracy for 6 datasets, 3 of which attain 100% classification accuracy. When using the NB and kNN classifiers, however, the recommended strategy yields classification accuracy that exceeds 91% for 6 datasets. Moreover, using the NB classifier, the proposed approach accomplishes the highest average classification accuracy, which is $4.88\%, 1.63\%, 7.64\%$, and 1.28% higher than that obtained using the mRMR, GR, IG, and Fisher procedures, respectively. Furthermore, the suggested technique, which is similar to the Fisher method, fulfills the highest average classification accuracy using the kNN classifier. It is important to note that the suggested technique using the SVM and NB classifiers outperforms others on the Breast dataset, a hard-classify dataset.

Figure 11 illustrates the average ranks of the five filter-based feature selection procedures using the three classifiers (the algorithm with the lowest rank is the one that performs the best, while the algorithm with the highest rank is the one that performs the worst).

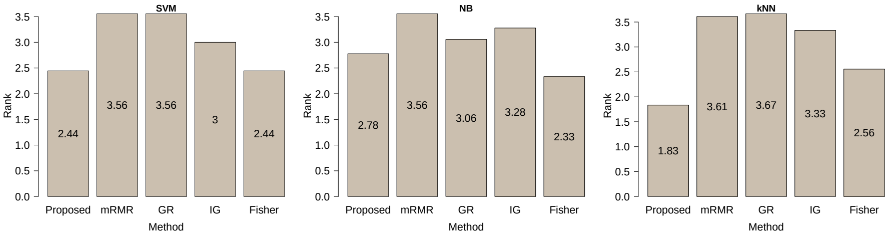

Fig. 11. Ranking of the proposed, mRMR, GR, IG, and Fisher algorithms by Friedman test.

The proposed method, which is similar to the Fisher algorithm, is ranked as the most significant by the SVM classifier, as can be seen in Figure 11. Furthermore, it has the second-best rank among the five filter-based approaches using the NB classifier. Moreover, the proposed method outperforms the other methodologies by achieving the lowest average rank when using the kNN classifier.

Comparison with no feature selection. The second experiment compares the classification performance of the proposed algorithm to that achieved without using any feature selection techniques (using all available features). Classification accuracy, sensitivity, precision, and F-measure are four performance metrics used for evaluation. The obtained results are displayed in Figures 12–14, respectively, using the SVM, NB, and kNN classifiers.

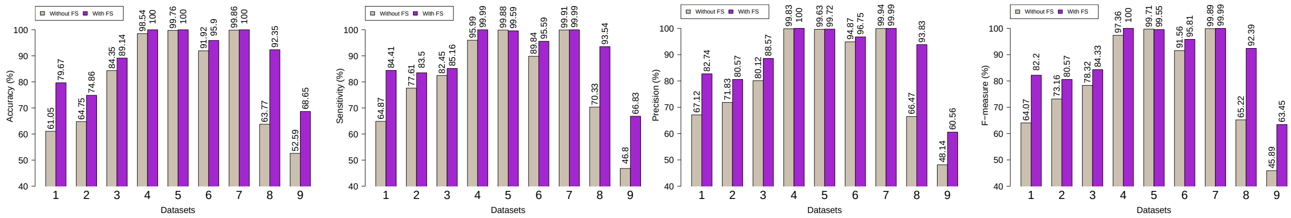

Fig. 12. Classification performances with and without feature selection using the SVM classifier.

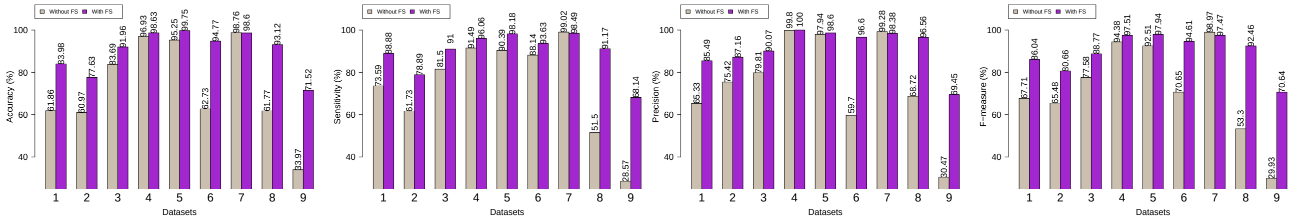

Fig. 13. Classification performances with and without feature selection using the NB classifier.

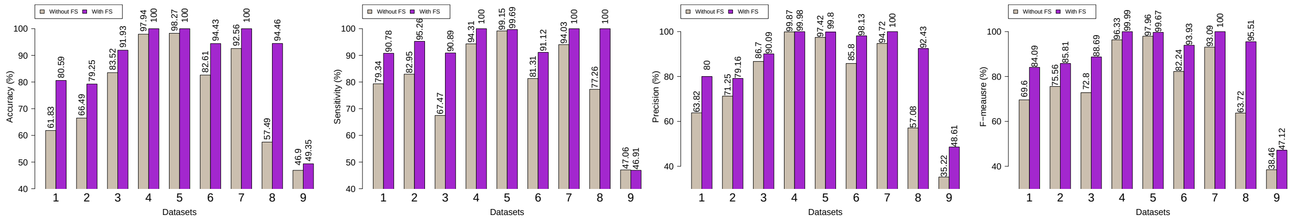

Fig. 14. Classification performances with and without feature selection using the kNN classifier.

In addition, Table 10 presents the numbers of the chosen genes as well as their relative proportions.

Based on the analysis of Figures 12–14 and Table 10, it is clear that the proposed algorithm employing the SVM, NB, and kNN classifiers improves classification accuracy, sensitivity, precision, and F-measure while also reducing the number of genes, thereby overcoming the dimensionality curse.

Table 10. Numbers of selected features and their relative proportions achieved by the proposed methodology using SVM, NB, and kNN classifiers.

Datasets	Original features	Selected features			Proportion (%)		
		SVM	NB	KNN	SVM	NB	KNN
Breast Cancer	4348	12	7	3	0.28	0.16	0.07
CNS	7129	5	33	25	0.07	0.46	0.35
Colon	2000	6	10	12	0.30	0.50	0.60
Leukemia	7129	33	4	32	0.46	0.06	0.45
Lymphoma	4026	30	21	29	0.75	0.52	0.72
Prostate	12600	16	11	4	0.13	0.09	0.03
SRBCT	2308	37	37	18	1.60	1.60	0.78
Westbc	7129	3	40	3	0.04	0.56	0.04
9_Tumors	5726	21	9	28	0.37	0.16	0.49

5. Conclusion

This work aimed to introduce a new ensemble filter approach based on PROMETHEE, it has the advantages of two filters MaCΨ-filter and PCRWG-filter. The proposed approach pre-selects a gene subset from the microarray dataset, which is characterized by high dimensions, numerous irrelevant genes, and a small sample size, for classification. The proposed approach is a novel multi-filter ensemble technique used to produce more compact gene subsets by integrating the outcomes of MaCΨ-filter and PCRWG-filter. Experiments were conducted using nine benchmark microarray datasets. The results showed that the proposed approach achieved competitive accuracies compared to four individual filters. Further research direction could examine the interaction between our approach and graph theoretic concept to further achieve a better classification performance on gene selection problem.

[1] Ang J. C., Mirzal A., Haron H., Hamed H. N. A. Upervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics. **13** (5), 971–989 (2015).

[2] Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks. **5** (4), 537–550 (1994).

[3] Alhaj T. A., Siraj M. M., Zainal A., Elshoush H. T., Elhaj F. Feature selection using information gain for improved structural-based alert correlation. PloS One. **11**, e0166017 (2016).

[4] Karegowda A. G., Manjunath A. S.,Jayaram M. A. Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management. **2** (2), 271–277 (2010).

[5] Sun L., Wang T., Ding W., Xu J., Lin Y. Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification. Information Sciences. **578**, 887–912 (2021).

[6] Javandel V., Vakilian M., Firuzi K. Multiple partial discharge sources separation using a method based on laplacian score and correlation coefficient techniques. Electric Power Systems Research. **210**, 108070 (2022).

[7] Ahakonye L. A. C., Nwakanma C. I., Lee J.-M., Kim D. S. SCADA intrusion detection scheme exploiting the fusion of modified decision tree and Chi-square feature selection. Internet of Things. **21**, 100676 (2023).

[8] Potharaju S. P., Sreedevi M. Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. Clinical Epidemiology and Global Health. **7** (2), 171–176 (2019).

[9] Yu L., Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. Proceedings of the 20th international conference on machine learning (ICML-03). 856–863 (2003).

[10] Shreem S. S., Abdullah S., Nazri M. Z. A., Alzaqebah M. Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection. Journal of Theoretical and Applied Information Technology. **46** (2), 1034–1039 (2012).

[11] Radovic M., Ghalwash M., Filipovic N., Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. **18**, 9 (2017).

[12] Chamlal H., Ouaderhman T., Aaboub F. A graph based preordonnances theoretic supervised feature selection in high dimensional data. Knowledge-Based Systems. **257**, 109899 (2022).

[13] Chamlal H., Ouaderhman T., Rebbah F. E. A hybrid feature selection approach for Microarray datasets using graph theoretic-based method. Information Sciences. **615**, 449–474 (2022).

[14] Chamlal H., Ouaderhman T., El Mourtji B. Feature selection in high dimensional data: A specific preordonnances-based memetic algorithm. Knowledge-Based Systems. **266**, 110420 (2023).

[15] Venkata Rao R., Patel B. K. Decision making in the manufacturing environment using an improved PROMETHEE method. International Journal of Production Research. **48** (16), 4665–4682 (2010).

[16] Vapnik V. The Nature of Statistical Learning Theory. Springer Science & Business Media, New York (1999).

[17] Rish I. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence. **3**, 41–46 (2001).

[18] Cover T., Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. **13**, 21 (1967).

[19] Ding C., Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology. **3** (2), 185–205 (2005).

[20] Demšar J. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research. **7**, 1–30 (2006).

Метод на основі фільтра PROMETHEE для даних експресії генів мікроматриці

Уадерман Т., Абуб Ф., Чамлал Х.

Кафедра математики та інформатики,
лабораторія фундаментальної та прикладної математики,
факультет наук Айн Чок, Університет Хасана II, Касабланка, Марокко

Набори даних експресії генів успішно застосовуються для різних цілей, включаючи класифікацію раку. Проблеми, з якими стикаються при розробці ефективних класифікаторів для наборів даних виразів, полягають у великій вимірності та перенавчанні. Відбір генів є ефективним і діючим методом подолання цих проблем і підвищення точності прогнозування класифікатора. Базуючись на PROMETHEE, ця стаття представляє ансамблевий підхід з декількома фільтрами шляхом інтеграції результатів двох потенційних фільтрів, а саме: MaCΨ-фільтра та PCRWG-фільтра для попереднього вибору найбільш інформативних генів. Були проведені експерименти на дев'яти наборах даних мікроматриці, щоб продемонструвати ефективність запропонованого методу.

Ключові слова: *фільтр; класифікація; відбір; дані мікроматриці; ПРОМЕТЕЙ.*