Massively Parallel Feature Selection Based on Ensemble of Filters and Multiple Robust Consensus Functions for Cancer Gene Identification

Anouar Boucheham and Mohamed Batouche

Abstract Currently, cancer prevails as a prime health matter worldwide. Selecting the appropriate biomarkers for early cancer detection might improve patient care and have often driven revolutions in medicine. Statistics and machine learning techniques have been broadly investigated for biomarker identification, especially feature selection where researchers try to identify the most distinguishing genes that can achieve better predictive performance of cancer subtypes. The robustness of the selected signature remains a crucial goal in personalized medicine. Ensemble and parallel feature selection are promising techniques to overcome this problem in which they have seen an increasing use in biomarker discovery. We focus in this chapter on the principal aspects of using ensemble feature selection in biomarker discovery. Furthermore, we propose a massively parallel meta-ensemble of filters (MPME-FS) to select a robust and parsimonious subset of genes. Two types of filters (ReliefF and Information Gain) are investigated in this study. The performances of the proposed approach in terms of robustness, classification power and the biological meaning of the selected signatures on five publicly available cancer datasets are explored. The results attest that the MPME-FS approach can effectively identify a small subset of biomarkers and improve both robustness and classification accuracy.

Keywords Bioinformatics · Biomarker discovery · Ensemble feature selection · Meta learning · ReliefF · Information gain · Cancer classification

A. Boucheham (✉) · M. Batouche
Computer Science Department, College of NTIC, Constantine 2 University,
MISC Laboratory, 25000 Constantine, Algeria
e-mail: anouar.boucheham@univ-constantine2.dz

M. Batouche
e-mail: mohamed.batouche@univ-constantine2.dz

K. Arai et al. (eds.), *Intelligent Systems in Science and Information 2014*,
Studies in Computational Intelligence 591, DOI 10.1007/978-3-319-14654-6_6

1 Introduction

Over the last two decades, the general rate of deaths to new cancer cases persists as high as 49 % overall. Thus, current bioinformatics efforts are focusing on biomarker discovery which is the key element of personalized medicine, where the genetic constitution is used to guide therapeutic approaches [1]. Therefore, the discovery of more effective cancer biomarkers is urgently needed, since the development and the effective use of biomarkers in clinical practice will certainly lead to tailor treatments for the disease in an individual [2].

The application of omics high-throughput technologies for cancer biomarker discovery is being rapidly expanded in current biomedical research, including DNA microarrays, Next Generation Sequencing (NGS) and MicroRNAs which are able to capture a substantial fraction of a cell state [3]. These technologies allow monitoring the expression levels of thousands of genes simultaneously in healthy and diseased cells, as well as are essential to biomarker discovery.

Gene expression data can effectively help to differentiate between cancer subtypes and then serve as an effective tool for diagnostic purposes in clinical practice. However, the identification of the smallest possible set of genes that could be used as biomarkers is a crucial problem in bioinformatics and personalized medicine. These genes must be the most informative for cancer prediction through supervised classification models [4]. The identification is generally referred to as a feature selection problem which is desirable to provide the features that contribute most to both classification and prediction [5]. It is a vital preprocessing step in data mining tasks, to reduce the effect of noise and improve the quality of data processing as well as considered to be one of current challenges in statistical machine learning for high-dimensional data.

A major challenge in the analysis of gene expression data is due to their sizes: a very small number of samples, of the order of tens, versus thousands of genes associated to all samples. This is commonly known as the "curse-of-dimensionality" which is also characterized by a large number of irrelevant, redundant and noisy genes that mislead or impede diagnosis efficiency [6]. Thus, only a fraction of genes contains useful biological interpretations and further gives a high accuracy for cancer diagnosis. Another challenge concerns the biological variations in real clinical tests which require the development of more stable feature selection methods [7]. In other words, selection of informative genes and an appropriate assessment of robustness, classification accuracy and biological meaning of the results are the most important matters in this field.

Ensemble-based learning is a robust and popular technique, due to the immense success of many ensemble methods in bioinformatics applications. It has the broad advantage of overcoming the curse-of-dimensionality in gene expression data, thereby offer higher accuracy and stability than conventional feature selection algorithm can achieve. Therefore, the use of ensemble methods to feature selection problem has been one of the recent growing trends. It consists of performing multiple diverse selectors with different subsamples, and then aggregates their results using a consensus function to obtain a final best subset of biomarkers [8]. Another benefit of applying ensemble feature selection, that it is naturally susceptible to parallelism, as well as we can easily undertake their parameters in parallel. The parallel

implementation of ensemble methods can certainly speedup the computational time of the selection and allow solving large-scale problems by involving multiprocessors to execute the different parts of the ensemble in parallel [9].

This chapter will focus on the different aspects of the application of ensemble feature selection methods to biomarker discovery from gene expression data. Furthermore, we propose a massively parallel meta-ensemble based feature selection method that can select robust and accurate biomarkers from DNA-microarrays datasets and can be generalized to several genomics studies. Two types of filter-based feature selection algorithms are investigated in this study: ReliefF and Information Gain. We also discuss the results in terms of robustness, classification power and the biological meaning of the selected signatures.

2 Application of Ensemble Feature Selection to Biomarker Discovery

In analogy with ensemble methods in supervised machine learning which combine multiple learned models to achieve high classification accuracy such as bagging and boosting [10]. Ensemble feature selection has received much attention recently. We mainly present here the different aspects to be considered in ensemble-based feature selection for biomarker discovery in which can help researchers to classify any method of them. The main critical problems in this category of methods are both the construction of diverse local selectors and the consensus function used to combine the different subsets of features [7, 11]. Therefore, the first aspect to be examined is the diversity design within the ensemble. This criterion divides ensemble feature selection methods into three classes:

- Ensemble based on data diversity: where we run the same selector with different subsamples generated from the original dataset [12, 13].
- Ensemble based on functional diversity: where different selectors are performed on the whole set of data (without sampling) [14].
- Ensemble based on data and functional diversity: here both data and functional diversity are combined in which multiple feature selection algorithms are performed on different subsamples.

Another aspect to be considered in ensemble feature selection methods is the representation used by the different selectors, since the notation of the results is not the same in all feature selection algorithms. This has a great impact on the consensus function to be used to aggregate the results [12]. Typically, we can observe three types of representations:

- Feature subset representation: subset containing only selected features (generally with different size)
- Feature ranking representation: subset of ranked features (a threshold is necessary for the selection)
- Feature weighting representation: a subset of pairs feature/weight which can easily converted to feature ranking representation.

Recent studies have focused on ensemble methods using wrapper-based selectors [15, 16]. It prompts us to consider the dependence of the selectors to any classifier as an important aspect in ensemble feature selection methods. This criterion influences the quality of solutions within the ensemble and the overall computational cost of the selection, as well as it divides ensemble feature selection methods into:

- Filter/ranking-based ensembles: they are simple, fast and independent of any classifier [14].
- Wrapper-based ensembles: they are very computationally intensive and have the risk of over-fitting due to high dimensionality of data as well as include the interaction between feature subset search and the mining algorithm. Moreover, they have the ability to take into account feature dependencies [17].
- Embedded-based ensembles: they use internal information of the classifier to perform selection and show a better computational complexity than wrapper methods [13].
- Hybrid-based ensembles: they are a combination of filter and wrapper methods which use the ranking information obtained using filters to guide the search in the optimization algorithms used by wrapper methods [16].

3 Massively Parallel Meta-Ensemble Feature Selection

Feature selection is an important preprocessing step in many machine learning applications including bioinformatics and computational biology, where it is generally used to find the smallest subset of features that extremely increases the performance of the classification model. In this section, we focus on ensemble of ensembles learning techniques which work by aggregating the outcomes of different ensembles into a final agreed decision through one or more consensus functions. The main objective is to attempt high performance of computer-aided diagnosis (CAD), by selecting a few genes with high predictive power and high sensibility to variations in real clinical tests. The selected biomarkers will be directly used by the CAD system for cancer diagnosis or others predictive goals.

For this purposes, a new parallel framework of feature selection is explored (see Fig. 1). In analogy with meta-ensemble models for supervised learning [10], the proposed approach is designed as an ensemble of ensembles of different selectors which perform selection in parallel through various ensembles and two consensus functions. In the following, we introduce our parallel framework for biomarker discovery in detail. We first, formulate both problem and representation of our solution under the proposed framework and then the general framework is explored including the parallel construction of the ranked lists as well as the related consensus functions.

Accordingly, biomarker discovery from gene expression is the problem of selecting subset of representative biomarkers from a large dataset. Given a set X of

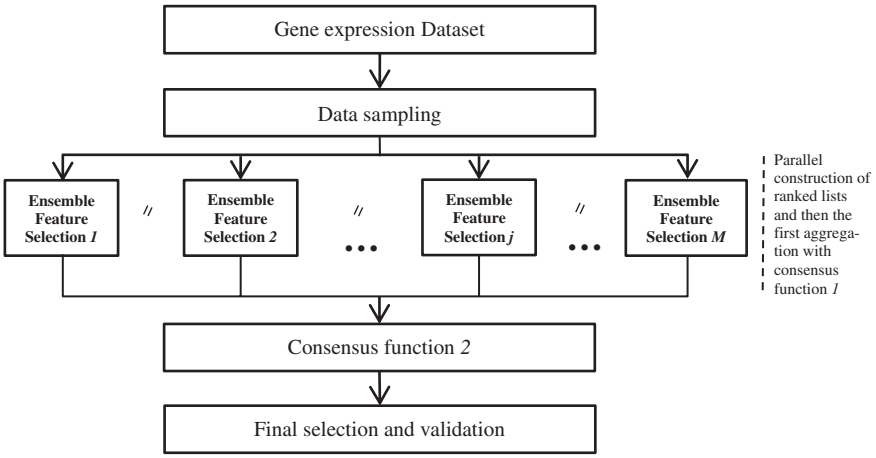

Fig. 1 Parallel model of the MPME-FS

K features with K very large, the problem consists in finding out the minimal subset $X_s^* \subset X$ that contains the more relevant yet non redundant features. Ensemble feature selection is a promising technique for addressing these complex structures of data and alleviates the problems of small sample size and high dimensionality [18].

The use of an ensemble of ensembles of filters leads to several subsets. Let us denote by X_{sj}^i the subset j of selected features using filter i. Therefore, two matters need to be addressed. The first one is related to the importance of each feature and the second one is related to the way subsets are aggregated to lead to the final subset of features. In order to properly deal with these two issues, we propose a two stage approach that uses ranking and consensus functions. At a first step various subsets of ranked lists of features are constructed using several filters then aggregation of these subsets is performed at two levels to form ensembles and then the meta-ensemble. More formally, the output of the first step can be represented as:

$$X_{sj}^i = \left\{ \left(f_{ij}^k, w_{ij}^k \right) \text{ where } i,j = 1 \ldots (N,M) \text{ and } k = 1 \ldots K \right\} \tag{1}$$

f_{ij}^k represents the rank of the feature k in the ensemble j using filter i. Its relevance is given by the weight w_{ij}^k. The global weight of feature k within the ensemble j is denoted as w_j^k. Three subsets of pairs of features and their weights are needed.

- The first is $Lbest_j$ that represents the local best features in each ensemble.

$$Lbest_j = \left\{ \left(k, w_j^k \right) \text{ where } j = 1 \ldots M \text{ and } k = 1 \ldots K \right\} \tag{2}$$

- This subset is the result of an aggregation process over $\mathrm{X}_{\mathrm{sj}}^{i=1\ldots N}$ subsets.

- The second is *Gbest*. It represents the global best features over the meta-ensemble. It is the result of an aggregation process over $Lbest_j$.
- The third is $Fbest = X_s^* \subset X$. It represents the final selected features given by MPME-FS method.

3.1 General Framework of MPME-FS for Biomarker Discovery

The general framework of MPME-FS consists of multiple ensembles of filters performed in parallel each of which employs a robust consensus function to select the best subset within each ensemble. The next step is to aggregate the outcome of all ensembles using a second consensus function and finally select features with higher scores given by all filters from all ensembles as shown in Fig. 1. The selection processes starts by the construction of M sub-samples $Si = 1 \ldots M$ from the whole dataset. Then, the parallel selection is initiated in all the M ensembles. At this stage, each ensemble j constructs N ranked lists X_{sj}^i by using filter i. To achieve the goal of both functional and data diversity when constructing the X_{sj}^i lists, we have used data partitioning with overlap allowing creating a reduced dataset to each $filter_i$.

Data perturbation involves generating subsamples by removing instances from the original datasets randomly. Knowing that, the overlap represents the percentage of samples belonging to the original dataset [11]. Subsequently, a consensus function within each ensemble is applied in order to aggregate these ranked lists X_{sj}^i and finally obtain the local best features in the ensemble j ($Lbest_j$) (see Fig. 2). Note that the construction of the ranked lists within each ensemble can be performed in parallel and through different filters (Information Gain, Gain ratio, Fisher Ratio, Symmetric uncertainty, ReliefF).

The following step consists of the aggregation of all local best features over the meta-ensemble, to construct the global best subset (Gbest) of features using the consensus function 2 alongside with the accumulation of scores associated

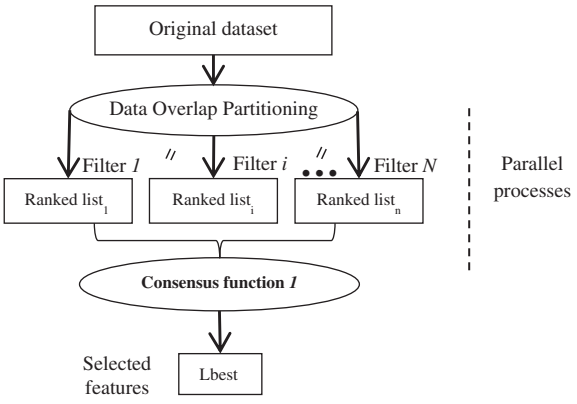

Fig. 2 Ensemble feature selection

to these features. Finally, we select the best ranked features which constitute the subset *Fbest*, from the subset *Gbest* ranked based on features' global weights

3.2 Consensus Functions

Recent work in biomarker identification has seen an increasing use of ensemble based feature selection due to their power to give higher accuracy and stability than a single algorithm can achieve. It can also dealing with small sample size and complex data structures. The key idea in ensemble methods is how to combine subsets of different selectors to lead to the final subset of features. This problem has received considerable attention in recent years [19]. Aggregation methods depend on the representation of the outcome of selectors which can be divided in three types: feature subset, feature ranking and feature weighting-score [11]. Based on these representations there exist many consensus functions, some of them include weighted voting, mean aggregation and threshold based aggregation for both rank and weighting-score representations, and counting the most frequently selected features for feature subset representation [12, 20].

Certainly, choosing the appropriate consensus function is a difficult task in ensemble methods. In our work, we use two consensus functions; the first one in the ensembles level and the second function is in the meta-ensemble level. Both are based on features ranking and their global weights which lead to a more robust and parsimonious final selection.

The first consensus function aggregates the ranked lists created by filters in the same ensemble. This function is inspired from both counting the most frequently selected features and weighted voting aggregation functions, but with hard selection by using the intersection over the entire ranked lists. For the intersection purpose, we use a threshold denoted by *TS1* in order to select only features belonging to *TS1* first ranked ones. Afterward, the weights of all selected features over the entire ensemble j denoted by $w_{i,j}^{k}$ are accumulated in order to obtain the global weights of selected features w_{j}^{k} over all ensembles [18]. More formally, the consensus function 1 can be described as follows:

$$\begin{cases} Lbest_j = \{\left(\boldsymbol{k}, \boldsymbol{w_j^k}\right)\} = \begin{cases} \bigcap_{i=1}^{N} \{\left(\boldsymbol{f_{i,j}^k}, \boldsymbol{w_{i,j}^k}\right)\} \\ \text{and } \boldsymbol{f_{i,j}^k} \leq TS1 \end{cases} \\ \text{where } \boldsymbol{w_j^k} = \sum_{i=1}^{N} w_{i,j}^{k} \end{cases}$$

By this way, we obtain the set *Lbestj* containing pairs of the best selected features with their global weights in the ensemble j $\{\left(k,\ w_{\mathrm{j}}^{\mathrm{k}}\right)\}$. This latter will be the input of the second consensus function in the meta-ensemble level, to construct the subset Gbest which contribute to the final selection.

The second consensus function consists primarily of aggregation the M subsets ($Lbest_j$) generated in the parallel previous step. The *Gbest* subset represents pairs

of features and their accumulated weights belonging to the union of the M local best subsets of features, which can be calculated as follows:

$$\begin{cases} Gbest = \left\{ \left(\boldsymbol{k}, \boldsymbol{bw^k} \right) \right\} = \bigcup_{j=1}^{M} Lbest_j \\ \text{where } \boldsymbol{bw^k} = \sum_{j=1}^{M} w_j^k \end{cases}$$

Finally, we select the best *TS2* ranked features from *Gbest* that represent the final selected features to be validated in the validation step. The pseudo-code of the whole process can be summarized as follows:

Algorithm: Massively Parallel Meta-Ensemble Feature Selection (**MPME-FS**).

Parameters:
N: ensemble size
M: meta-ensemble size
TS1: threshold of intersection
TS2: percentage of best selected features in meta-ensemble
Overlap: overlap of data sampling

Input:
D: dataset with *K* features
L: sample labels in *D*

Output:
Fbest: final best selected features

Parameters initialization:
Gbest, **Lbest** =Ø; ***Filter***= filter_i;

Parallel Meta ensemble feature selection process:

1: For (each ensemble_j / j=1 ….M) **do in parallel**

// determining Lbest_j for each ensemble_j

2: For (i=1…N) **do in parallel**
3: S_i = sampling (D, *Overlap*)
4: $(k, w_{i,j}^k)$ = *Filter* (S_i, L)
5: End For in parallel

// obtaining the rank $f_{i,j}^k$ of each feature in the ensemble_j

6: $\{(f_{i,j}^k, w_{i,j}^k)\}$ = Sort ($\{(k, w_{i,j}^k)\}$)

// aggregating results of all filters within the ensemble_j

7: ***Lbest***$_j$ = $\{(k, w_j^k)\}$ = **Consensus function 1** ($\{(f_{i,j}^k, w_{i,j}^k)\}$)

8: End For in parallel

// aggregating all *Lbest*$_j$ subsets

9: ***Gbest*** = **Consensus function 2** (***Lbest***$_j$)

// final selection

10: Sort (***Gbest***) // sort k based on bw^k
11: ***Fbest*** = select *TS2* best features from ***Gbest***

---

4 Experiments and Discussions

In the following sections, the analysis of classification performances, robustness and biological interpretation of the MPME-FS method on large feature and small sample size microarrays are presented. First, the data sets and the experimental settings used in this analysis are briefly described. Second, we analyze the classification performances in terms of accuracy, sensitivity and specificity using different classifiers. After that, we study the robustness of the selected signatures. Finally, we perform a biological interpretation of the selected genes.

4.1 Datasets and Experiment Setting

All experiments were conducted using MATLAB®'s Parallel Computing Toolbox (PCT). The proposed MPME-FS was evaluated by means of five publicly available DNA microarray datasets which can be divided into binary and multiclass types. The binary datasets are the most prominent and can separate healthy patients from cancer patients, while multiclass datasets are used to differentiate the various types of cancers based on gene expressions. Therefore, the datasets were collected from both Kent Ridge bio-medical data repository[1] and Gene Expression Model Selector, from Vanderbilt University[2]. The main datasets characteristics are shown in Table 1.

To assess the performances of our parallel meta-ensemble feature selection method, we use in the experiments two well-known and successful filters: information gain and ReliefF. Based on an empirical evaluation using different settings of the proposed method, the best parameters setting of MPME-FS which is adopted in this study is depicted in Table 2.

4.2 Classification Accuracy Analysis

The first experiment is devoted to assess the performance of the MPME-FS in terms of accuracy, sensitivity, specificity and the number of selected biomarkers using 10-fold cross validation technique. The latter is a common choice in the specialized literature [21], which splits the whole set of data into many subsets to evaluate the goodness of the selected signature.

To achieve high level evaluation of the classification ability of the selected genes, we first use different classifiers separately (SVM, KNN, ANN). Then we have employed an ensemble of different classifiers (SVM, KNN and ANN) with

[1] http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html.

[2] http://www.gems-system.org.

Table 1 Characteristics of the different datasets used for evaluation

Dataset	#Features	#Classes	#Samples
Ovarian	15,154	2	253
Leukemia	7,129	2	72
DLBCL	5,469	2	77
Colon	2,000	2	62
SRBCT	2,308	4	83

Table 2 MPME-FS parameters setting

Filters	InfoGain, ReliefF
Ensemble size	10
Meta-ensemble size	100
TS1	150
TS2	30
Overlap	80
K value in ReliefF	10

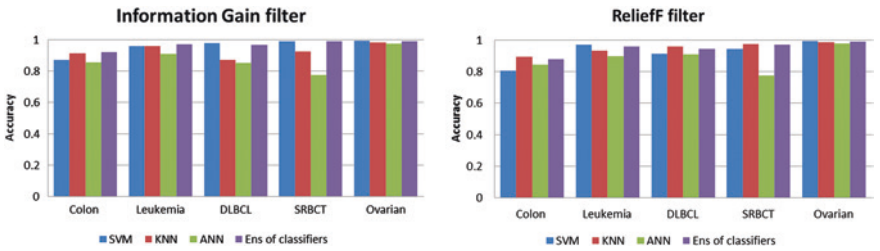

Fig. 3 Average 10 cross-validation classification accuracy using: SVM, KNN, ANN and an ensemble of different classifiers over the five datasets

majority voting as consensus function. Accordingly, the results indicated in Fig. 3 represent the average accuracies of MPME-FS given by SVM, KNN, ANN and the ensemble of classifiers described above. For comparison reasons, we have used in this experiment both Information Gain and ReliefF filters to perform selection by the proposed approach. Form this figure we observe that the MPME-FS performs better using information Gain filter than ReliefF filter in the five datasets.

A second observation which can be made is that the ensemble of classifier gives higher accuracy than single classifiers in almost cases that are not surprising since it combines the efforts of the three classifiers. Furthermore, Fig. 4 shows boxplots of MPME-FS using SVM classifier over thirty runs for both Information Gain and ReliefF (Fig. 4a, b successively). As desired, the performance variance between runs reaches an almost completely stable result through the two filters (among 0.001 and 0.04).

We also provide in Table 3 the average performance of MPME-FS using the different classifiers of both information Gain and ReliefF filters. In the last column

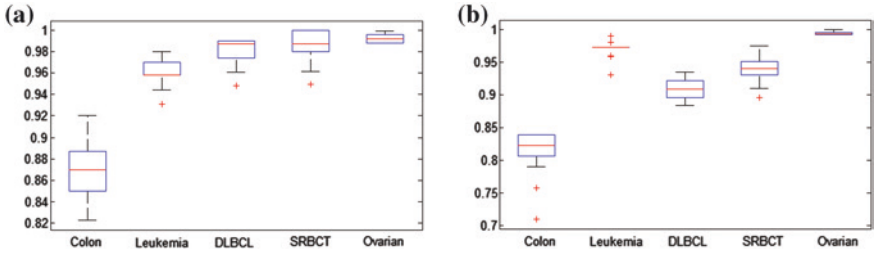

Fig. 4 Boxplots of MPME-FS method on colon, leukemia, DLBCL, SRBCT and ovarian datasets across 30 runs. **a** Informain gain filter and **b** ReliefF filter

Table 3 Average classification results in terms of sensitivity (sensi), specificity (speci) and the number of selected biomarkers (# genes) of MPME-FS using both information gain and ReliefF filters over the five datasets

		SVM		KNN		ANN		Ensemble of classifiers		# genes
		Sensi	*Speci*	*Sensi*	*Speci*	*Sensi*	*Speci*	*Sensi*	*Speci*	
InfoGain	Colon	0.875	0.804	0.925	0.907	0.871	0.831	0.914	0.909	30
	Leukemia	0.957	1	0.978	0.96	0.934	0.92	0.953	1	31
	DLBCL	0.965	1	0.827	1	0.924	0.63	0.952	1	27
	SRBCT	1	0.94	0.896	1	0.89	0.92	1	0.963	32
	Ovarian	0.998	1	0.993	0.967	1	0.978	1	0.988	39
Average		**0.959**	**0.948**	**0.923**	**0.966**	**0.923**	**0.855**	**0.963**	**0.972**	**31**
ReliefF	Colon	0.875	0.81	0.925	0.863	0.9	0.818	0.89	0.863	31
	Leukemia	0.878	0.96	1	0.84	0.872	0.88	1	0.96	33
	DLBCL	0.931	0.947	0.965	0.947	0.948	0.842	0.948	0.947	27
	SRBCT	0.931	0.944	0.965	1	0.931	0.925	0.965	0.981	33
	Ovarian	1	1	1	0.978	1	0.978	1	0.989	41
Average		**0.923**	**0.932**	**0.971**	**0.925**	**0.9302**	**0.888**	**0.960**	**0.948**	**33**

the number of selected biomarkers on the five datasets is shown. We observe that both sensitivity and specificity of our selection are convergent among the different classifiers. The previous experiments were performed on 30 independent runs to have statistically meaningful conclusions as our approach is stochastic.

4.3 Robustness Analysis

We explore and discuss in the present study the robustness of the selected signature by the MPME-FS approach. Therefore, we assess the similarity between the outputs of different independent executions of our method. The global stability is

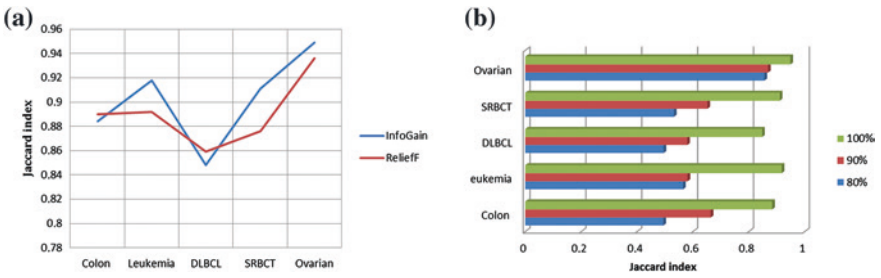

Fig. 5 Average robustness results of the MPME-FS in term of Jaccard index over 20 independent runs on the five datasets. **a** Information Gain filter versus ReliefF filter **b** Jaccard index versus perturbation rate of subsampling between the independent runs (80, 90 and 100 %)

defined as the average over all pairwise similarity comparisons between the different feature selectors as follows [12]:

$$S_{tot} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S(f_i, f_j)}{k(k-1)} \tag{3}$$

where f_i represents the outcome of the feature selection method applied to subsample$_i$ ($1 \leq i \leq 20$), and $S(f_i, f_j)$ represents a similarity measure between f_i and f_j. Mainly, for feature subsets selection (as in our case), we use the Jaccard index (JI) which can be calculated as follows:

$$S(f_i,f_j) = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} \tag{4}$$

A set of experiment assess the overall stability of the selected signature on the five datasets using both Information Gain and ReliefF filters which is shown in the Fig. 5a. Results in term of Jaccard index show that the MPME-FS performed using information gain is generally more robust over the most datasets. To provide a better robustness analysis, we assess the effect of data perturbation rate when creating subsamples on the stability of the signature. In this experiment we use Information Gain as filter over the five datasets of which the results can be seen in Fig. 5b, which indicates that the robustness decreases as the perturbation rate is decreased.

4.4 Biological Interpretation of the Results

In this section, we address biological analysis of the selected biomarkers. We focus in this experiment on the analysis of the selected biomarkers from Colon and Leukemia datasets which are widely studied in the literature. Accordingly, Tables 4 and 5 list and describe the top thirty ranked genes over 30 independent

Table 4 Description of the top thirty selected genes from colon dataset, with a complete frequency level (freq = 30) over 30 independent runs

Gene index	Accession number	Gene description
66	T71025	3′ UTR 1 84103 Human (HUMAN)
1423	J02854	Gene 1 "MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN); contains element TAR1 repetitive element"
1414	R64115	3′ UTR 2a 139618 ADENOSYLHOMOCYSTEINASE (Homo sapiens)
137	D25217	Gene 1 "Human mRNA (KIAA0027) for ORF, partial cds"
138	M26697	Gene 1 "Human nucleolar protein (B23) mRNA, complete cds"
241	M36981	Gene 1 "Human putative NDP kinase (nm23-H2S) mRNA, complete cds"
245	M76378	Gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6"
249	M63391	Gene 1 "Human desmin gene, complete cds"
267	M76378	Gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6"
1843	H06524	3′ UTR 1 44386 "GELSOLIN PRECURSOR, PLASMA (HUMAN)"
286	H64489	3′ UTR 2a 238846 LEUKOCYTE ANTIGEN CD37 (Homo sapiens)
365	X14958	Gene 1 Human hmgI mRNA for high mobility group protein Y
377	Z50753	Gene 1 H.sapiens mRNA for GCAP-II/uroguanylin precursor
1960	D59253	Gene 1 Human mRNA for NCBP interacting protein 1
493	R87126	3′ UTR 2a 197371 "MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)"
513	M22382	Gene 1 MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN)
625	X12671	Gene 1 Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
739	X12369	Gene 1 "TROPOMYOSIN ALPHA CHAIN, SMOOTH MUSCLE (HUMAN)"
897	H43887	3′ UTR 2a 183264 COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
765	M76378	Gene 1 "Human cysteine-rich protein (CRP) gene, exons 5 and 6"
780	H40095	3′ UTR 1 175181 MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN)
812	Z49269	Gene 1 H.sapiens gene for chemokine HCC-1
964	T86473	3′ UTR 1 114645 NUCLEOSIDE DIPHOSPHATE KINASE A (HUMAN)
1042	R36977	3′ UTR 1 26045 P03001 TRANSCRIPTION FACTOR IIIA
1411	H77597	3′ UTR 1 214162 H.sapiens mRNA for metallothionein (HUMAN)
1494	X86693	Gene 1 H.sapiens mRNA for hevin like protein
1582	X63629	Gene 1 H.sapiens mRNA for p cadherin
1635	M36634	Gene 1 "Human vasoactive intestinal peptide (VIP) mRNA, complete cds"
1771	J05032	Gene 1 "Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds"
1263	T40454	3′ UTR 2a 60221 ANTIGENIC SURFACE DETERMINANT PROTEIN OA3 PRECURSOR (Homo sapiens)

Table 5 Description of the top thirty selected genes from Leukemia dataset, with a complete frequency level (freq = 30) over 30 independent runs

Gene index	Accession number	Gene description
758	**D88270_at**	GB DEF = (lambda) DNA for immunoglobin light chain
760	D88422_at	CYSTATIN A
1144	**J05243_at**	SPTAN1 Spectrin, alpha, nan-erythrocytic 1 (alpha-fodrin)
1630	L47738_at	Inducible protein mRNA
1685	**M11722_at**	Terminal transferase mRNA
1834	**M23197_at**	CD33 CD33 antigen (differentiation antigen)
1882	**M27891_at**	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
1902	M29474_at	Recombination activating protein (RAG-1) gene
2121	**M63138_at**	CTSD Cathepsin D (lysosomal aspartyl protease)
2128	M63379_at	CLU Clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)
2288	**M76559_at**	Neuronal DHP-sensitive, voltage-dependent, calcium channel alpha-2b subunit mRNA
2354	**M92287_at**	CCND3 Cyclin D3
2363	M93056_at	LEUKOCYTE ELASTASE INHIBITOR
2402	M96326_rna1_at	Azurocidin gene
2642	**U05259_rna1_at**	MB-1 gene
3252	**U46499_at**	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
4107	**X07743_at**	PLECKSTRIN
4196	**X17042_at**	PRG1 Proteoglycan 1, secretory granule
4328	**X59417_at**	PROTEASOME IOTA CHAIN
4366	**X61587_at**	ARHG Ras homolog gene family, member G (rho G)
4377	**U46499_at**	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
4847	**X95735_at**	Zyxin
5171	Z49194_at	OBF-1 mRNA for octamer binding factor 1
5501	**Z15115_at**	TOP2B Topoisomerase (DNA) II beta (180kD)
6041	**L09209_s_at**	APLP2 Amyloid beta (A4) precursor-like protein 2
6281	**M31211_s_at**	MYL1 Myosin light chain (alkali)
6855	**M31523_at**	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
1909	M29696_at	IL7R Interleukin 7 receptor
1953	M33195_at	Fc-epsilon-receptor gamma-chain mRNA
2335	M89957_at	IGB Immunoglobulin-associated beta (B29)

runs which have a complete frequency level (freq = 30) from Colon and Leukemia datasets successively. Furthermore, the selected genes are considered informative in most well-known methods in the literature. Specially, in Leukemia dataset which has been widely studied in this area.

As a result, genes selected from Leukemia dataset listed in boldface in Table 5, were also selected among the top 25 most relevant genes by Wu et al. [4] and

they are considered useful to discriminate between the two class label AML and ALL. Moreover, eight out of these thirty top genes selected by our method, i.e., M23197_at, M27891_at, U05259_rna1_at, U46499_at, X95735_at, L09209_s_at, M31523_at and M89957_at were deemed as relevant by Zhu et al. [21]. The selected genes can now be validated by biologists through clinical trials. We expect these discoveries may offer useful information for biologists and medical experts.

5 Conclusion

In summary, we considered in this chapter the application of ensemble feature selection methods to biomarker identification. Indeed, the most reviewed ensemble feature selection methods attest that this technique is a promising direction for more stable and accurate selection in cancer gene identification. We have also proposed a massively parallel approach based on meta-ensemble of filters for biomarker discovery from high dimensional data. The MPME-FS is different from other ensemble feature selection methods since it performs a parallel selection in two steps: the first one within each ensemble by the aggregation of results of different selectors, the second step is the aggregation of the outcomes of all ensembles using a second consensus function. The final selected biomarkers are employed to construct a classification model that will be used as an effective tool to handle patients and diagnose cancer subclasses.

In addition, the proposed MPME-FS is very fast and is computationally efficient as it is massively parallel and no learning algorithm is used in the selection process. Instead, we have employed filter model which is usually exploited when the number of features becomes very large especially for high dimensional data. Clearly, the MPME-FS can be performed using any ranking based feature selection algorithm and applied to any feature selection problem.

The experiments over five DNA microarrays datasets revealed that good results can be achieved through MPME-FS in terms of classification performance and robustness. Biological analysis of the results shows that MPME-FS provides the selection of highly informative genes which have biological meanings and are also selected by the other approaches.

References

1. Zhang, X., et al.: Integrative omics technologies in cancer biomarker discovery. Omics Technol. Cancer Biomark. Discov. **129** (2011)
2. Nair, M., Sandhu, S.S., Sharma, A.K.: Prognostic and predictive biomarkers in cancer. Curr. Cancer Drug Targets (2014)
3. Mäbert, K., Cojoc, M., Peitzsch, C., Kurth, I., Souchelnytskyi, S., Dubrovska, A.: Cancer biomarker discovery: current status and future perspectives. Int. J. Radiat. Biol. (0), 1–48 (2014)

4. Wu, M.Y., Dai, D.Q., Shi, Y., Yan, H., Zhang, X.F.: Biomarker identification and cancer classification based on microarray data using laplace naive bayes model with mean shrinkage. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **9**(6), 1649–1662 (2012)
5. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowl. Inf. Syst. **34**(3), 483–519 (2013)
6. Bolón-Canedo, V., Sánchez-Maroño, N., et al.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014)
7. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Comput. Biol. Chem. **34**(4), 215–225 (2010)
8. Guan, D., Yuan, W., Lee, Y.K., Najeebullah, K., Rasel, M.K.: A review of ensemble learning based feature selection. IETE Tech. Rev. **31**(3), 190–198 (2014)
9. Upadhyaya, S.R.: Parallel approaches to machine learning—a comprehensive survey. J. Parallel Distrib. Comput. **73**(3), 284–292 (2013)
10. Yang, P., Hwa Yang, Y., B Zhou, B., Y Zomaya, A.: A review of ensemble methods in bioinformatics. Curr. Bioinf. **5**(4), 296–308 (2010)
11. Awada, W., Khoshgoftaar, T.M., et al.: A review of the stability of feature selection techniques for bioinformatics data. In: Information Reuse and Integration (IRI), 13th International Conference, 356–363 (2012)
12. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Machine Learning and Knowledge Discovery in Databases, pp. 313–325. Springer, Berlin (2008)
13. Abeel, T., Helleputte, T., et al.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics **26**(3), 392–398 (2010)
14. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Data classification using an ensemble of filters. Neurocomputing **135**, 13–20 (2014)
15. Yang, P., Liu, W., Zhou, B. B., Chawla, S., Zomaya, A.Y.: Ensemble-based wrapper methods for feature selection and class imbalance learning. In: Advances in Knowledge Discovery and Data Mining, pp. 544–555. Springer, Berlin (2013)
16. Xu, J., Sun, L., Gao, Y., Xu, T.: An ensemble feature selection technique for cancer recognition. Bio-Med. Mater. Eng. **24**(1), 1001–1008 (2014)
17. Ghorai, S., et al.: Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM Trans. Comput. Biol. Bioinf. **8**(3), 659–671 (2011)
18. Boucheham, A., Batouche, M.: Robust biomarker discovery for cancer diagnosis based on meta-ensemble feature selection. In: The Proceedings of Science and Information Conference, IEEE, pp. 452–460 (2014). ISBN: 978-0-9893193-1-7
19. Boulesteix, A.L., Slawski, M.: Stability and aggregation of ranked gene lists. Briefings Bioinf. **10**(5), 556–568 (2009)
20. Haury, A.C., Gestraud, P., Vert, J.P.: The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS ONE **6**(12), e28210 (2011)
21. Zhu, Z., Ong, Y.S., et al.: Identification of full and partial class relevant genes. Comput. Biol. Bioinf. IEEE/ACM Trans. **7**(2), 263–277 (2010)