Adv Data Anal Classif (2018) 12:937–952
https://doi.org/10.1007/s11634-017-0285-y

REGULAR ARTICLE

Ensemble feature selection for high dimensional data: a new method and a comparative study

Afef Ben Brahim[1] · **Mohamed Limam**[2]

Received: 15 January 2015 / Revised: 30 October 2016 / Accepted: 17 April 2017 /
Published online: 24 April 2017

Abstract The curse of dimensionality is based on the fact that high dimensional data is often difficult to work with. A large number of features can increase the noise of the data and thus the error of a learning algorithm. Feature selection is a solution for such problems where there is a need to reduce the data dimensionality. Different feature selection algorithms may yield feature subsets that can be considered local optima in the space of feature subsets. Ensemble feature selection combines independent feature subsets and might give a better approximation to the optimal subset of features. We propose an ensemble feature selection approach based on feature selectors' reliability assessment. It aims at providing a unique and stable feature selection without ignoring the predictive accuracy aspect. A classification algorithm is used as an evaluator to assign a confidence to features selected by ensemble members based on their associated classification performance. We compare our proposed approach to several existing techniques and to individual feature selection algorithms. Results show that our approach often improves classification performance and feature selection stability for high dimensional data sets.

Keywords Feature selection · Ensemble methods · Classification · Stability · High dimensionality

Mathematics Subject Classification 41A05 · 41A10 · 65D05 · 65D17

✉ Afef Ben Brahim
afef.benbrahim@yahoo.fr

Mohamed Limam
mohamed.limam@isg.rnu.tn

1 Université de Tunis, Tunis Business School, LARODEC, BP 65, 2059 BIR El Kassaa, Tunisia

2 Dhofar University, Salalah, Sultanate of Oman

1 Introduction

The rapid technological developments in different life domains increase the amounts of data at an unprecedented speed. This may appear useful for the decision making process, however it is not the case when this increase concerns dimensions of data. Examples of such data are measurements arising in face recognition from digitized images, spam email identification, diagnostic tasks in medicine and genetic engineering, recognition tasks in biology, etc. In microarray data analysis for example, each sample involves the measurements of tens of thousands of variables corresponding to the expression of tens of thousands of genes measurable with microarray technology. Unfortunately, existing machine learning methods are not designed to handle such data setting, because the ability to build models with scientific validity is negatively impacted by an increasing ratio between the number of variables and the sample size (Kohane et al. 2003). This phenomenon is known as the curse of dimensionality. A large number of features can increase the noise of the data and thus the error of a learning algorithm. Feature selection is a solution for such problems. It reduces data dimensionality by removing irrelevant and redundant features. There are three supervised feature selection categories namely, wrappers, filters and embedded methods. Many reviews of these methods are found in the literature. Guyon and Elisseff (2003), Saeys et al. (2007) are examples of such good reviews. Filters select subsets of features as a pre-processing step, independently of the chosen predictor. Wrapper and embedded methods, on the other hand, generally use a specific learning algorithm to evaluate a specific subset of features. Different feature selection algorithms will choose different feature subsets. We may not say that a resulting subset is better than the others but rather that all the obtained subsets are the best subsets among the whole feature space. To deal with this issue, we naturally think of ensemble learning (Dietterich 2000) as a way to combine independent feature subsets, obtained by varying data or base learners, in order to get a robust feature subset in terms of classification performance but also stability of the selection. The fusion of different feature selectors is a step to generate a new feature set from the individual selected sets of features. There are two possible alternatives to combine the results of multiple feature selection algorithms for classification problems which have been proposed in the literature. These two alternatives are based on two aggregation levels, classifier level and selector level. In the first aggregation level, different feature subsets are generated and used for constructing an ensemble of accurate and diverse base classifiers. Classifiers' outputs are then combined to obtain the final classification results. The second aggregation level finds a consensus between the results obtained by several feature selection methods in order to obtain a unique feature subset before the classification process. Since feature selection stability is as important as classification accuracy, we are interested on having a single and combined feature subset. Thus, we focus on promoting ensemble feature selection at the selector aggregation level. Hence, we propose an ensemble feature selection approach based on a robust feature aggregation technique to combine the feature selection ensemble.

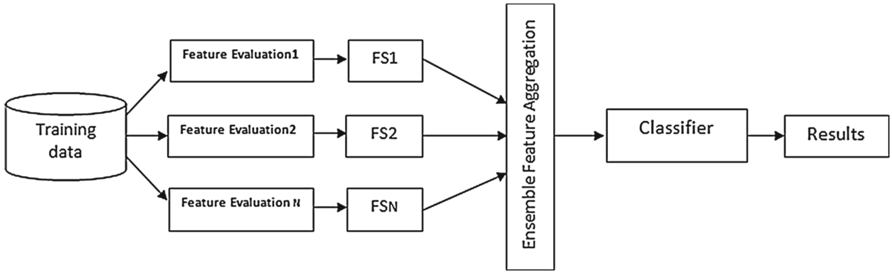

Fig. 1 Ensemble feature selection based selectors aggregation

2 Ensemble feature selection

The concept of ensemble feature selection based feature selectors' aggregation was introduced by Saeys et al. (2008). Similar to the case of supervised learning (Dietterich 2000), ensemble techniques might be used to improve the robustness of feature selection techniques. Different feature selection algorithms may yield feature subsets that can be considered local optima in the space of feature subsets. Ensemble feature selection could help in alleviating this problem by aggregating the outputs of several feature selectors. This concept was specially applied for high dimensional data with few samples as discussed by Saeys et al. (2008), Schowe and Morik (2011). Ensemble concept for feature selection can be also in the form of parallel application of multiple feature selection algorithms. Mitchell et al. (2014) proposed a parallel implementation of the bootstrap resampling step and combination of results of rank product method for feature selection for the identication of differentially expressed genes. Similar to the construction of ensemble models for supervised learning, there are two essential steps in creating a feature selection ensemble. The first step involves creating a set of different feature selectors, each providing an output, while the second step aggregates the results of the single models. Figure 1 illustrates the process of ensemble feature selection based selector aggregation.

2.1 Ensemble construction

In ensemble learning, a key point to obtain a good ensemble feature selection is to generate a diverse set of feature selections. There are two efficient ways for this purpose: the first uses the same type of base learner with different samples of data (homogeneous ensembles) and the second is based on different types of base learners trained on the same data (heterogeneous ensembles). A third alternative to achieve diversity would be to vary the data and the base learners at the same time and construct heterogeneous ensembles trained on different training sets.

2.2 Ensemble aggregation: feature selector level

To combine the resulting feature lists from the ensemble into a single decision for each feature, there exist simple aggregation techniques and other more complicated ones. Often, these techniques are used to aggregate either feature weights or ranks. We introduce some of these aggregation techniques:

Weighted mean aggregation (WMA) This method uses the weights of all the features obtained by the different selected subsets then for each feature the weights mean is calculated and features with the highest scores are selected (Abeel et al. 2010).

Complete linear aggregation (CLA) This method uses the complete ranking of all the features then the ranks over all ranking lists are summed for each feature. The best features are those with the lowest summed ranks (Abeel et al. 2010).

Robust RankAggregate (RRA) This method, proposed by Kolde et al. (2012), detects features that are ranked consistently better than expected under the null hypothesis of uncorrelated inputs and assigns a significance score for each feature. As a result, a p value is assigned for all items, showing how good it is positioned in the ranked lists than what is expected by chance.

Feature occurrence frequency (OFA) It obtains the final feature selection by calculating the number of occurrences of each feature over all lists and ranking them based on their occurrence frequency. This ranking technique favors features appearing in the maximum number of candidate feature subsets.

Classification accuracy based aggregation (CAA) Chan et al. (2008) proposed this method that assigns a score to each feature in the different lists as the sum of accuracies for all classifiers that include that feature. Such a scoring scheme favors the features that lead to more accurate classification but it is considered simple.

We propose a sophisticated and robust aggregation method to optimize classification accuracy and stability of feature selection based on features reliability assessment. The proposed approach is detailed in the following.

3 Robust ensemble feature selection based on reliability assessment

In ensemble feature selection, different feature selection algorithms are built based on optimizing different relevance criteria. Thus, they will have different biases and may produce different results. Okun (2011) mentioned that despite such a difference, if the same gene appears in multiple selected feature subsets obtained by different algorithms, and produce accurate classifiers, it is indeed important. We propose a robust feature selection aggregation technique based on this idea. The proposed ensemble feature selection framework consists of two steps. The first is the ensemble creation and the second is the ensemble outputs aggregation. These two important steps of the proposed method are detailed in the following.

3.1 Ensemble construction

The generation of diverse feature subsets can be achieved by constructing homogeneous ensembles where the same component learner is applied to different sample subsets or by constructing heterogeneous ensembles referring to those in which the component learners are different from each other.

3.1.1 Homogeneous ensembles

Starting from a particular training set, our aim is to generate a diverse set of feature selections. To generate diversity in the selection, the feature selection method is run on different training sub-samples. To this end, we make use of the bootstrapping method (Kohavi 1995), a well-established technique in statistics to reduce variance. We generate 30 bootstrap samples with replacement from the training data. This ensemble size is fixed after running experiments and its choice is heuristic based on the recorded classification performance and stability results. We apply a filter to each of the bootstrap samples to obtain a diverse set of feature rankings. In our experiments, we use three filters and thus we get three settings for homogeneous ensembles, each one using a filter. The filters are the same used in the construction of heterogeneous ensembles described below.

3.1.2 Heterogeneous ensembles

The basic idea in the construction of heterogeneous ensembles is to leverage on the strengths of different algorithms to obtain robust feature subsets. Consider a dataset $\mathbf{D}S = (x_i, \ldots, x_m), x_i = (x_i^1, \ldots, x_i^d)$ with m instances and d features. An heterogeneous ensemble of feature selection algorithms $(H_1, \ldots, H_K)$ is applied to DS resulting on K feature subsets $(F_1, \ldots, F_K)$ each one containing n selected features $\mathbf{F}_k = (f_{k,1}, \ldots, f_{k,n})$. For high dimensional data, filters are usually chosen as long as they are computationally efficient, fast and independent of the classification algorithm. Thus, we choose three popular and successful filters which are based on different selection criteria to create an ensemble of three selectors. These algorithms are detailed below.

Relief algorithm is a ranker proposed by Kira and Rendell (1992). It assigns a relevance weight to each feature to denote the relevance of the feature to the target concept. For each feature, it samples instances randomly from the training set and updates the relevance values based on the difference between the selected instance and the two nearest instances of the same and opposite class. Then, the feature is scored as the sum of weighted differences in the different class and the same class.

The minimum-Redundancy-Maximum-Relevance (mRmR) method proposed by Peng et al. (2005) is a mutual information based method. It selects features according to the maximal statistical dependency criterion. The mRMR method selects a feature subset that has the highest relevance with the target class, subject to the constraint that selected features are mutually as dissimilar to each other as possible.

The t test (Gosset 1908) prefers features with a maximal difference of mean value between groups and a minimal variability within each group. The t test is used in the form that defines the score of a feature as the ratio of the difference between its mean values for each of the two classes and the standard deviation.

3.1.3 Heterogeneous ensembles with varying data

This type of ensembles is based on combining both data variation and algorithm variation. We first generate 10 different bootstrap samples from the data as done for homogeneous ensembles, then for each bootstrap sample we apply the three different algorithms described above as for the heterogeneous ensembles. Thus, 30 different sets of features are obtained.

For each of the three described settings, we select a feature subset of best features from each of the obtained ranking lists. Filter methods give as output all the input features ranked according to their score so we don't have any indication about the feature set size required to have a good classification performance. A way to approximate the best solution would be to evaluate many feature set cardinalities with a classification algorithm and to keep the cardinality that gives the best classification performance. In this pre-processing phase, we choose a cardinality of 1% of the initial features number for the candidate subsets obtained in this step.

3.2 Ensemble aggregation based on reliability assessment

The choice of the technique to use for the aggregation step is an important decision for ensemble feature selection We propose a robust feature aggregation technique which objective is to improve robustness of the results, i.e. classification performance but also feature selection stability. To this end, the proposed ensemble aggregation technique uses the classification performance obtained with different feature subsets to guide the selection of features corresponding to high accuracies. The feature selector's confidence and their conflict with other selectors in the ensemble are measured in order to assign a reliability factor guiding the final feature selection. Therefore, the proposed reliability assessment based aggregation (RAA) technique involves two steps. The first one is the features' confidence calculation based on their weights and associated classification error. The second one is the reliability assessment and decision making.

3.2.1 Confidence calculation

We note that the trained feature selectors ensemble resulted in K feature subsets. A classifier is trained on each newly obtained training set containing only the feature subset obtained by each feature selector. The overall accuracies of the K classifiers by tenfold cross validation (CV) are determined. Each classifier is used here to evaluate an individual feature subset. It assigns a confidence level according to the classification performance obtained with the projection of that feature subset. Any classification algorithm could be used but it is preferable to choose a simple classifier as we are still in a preprocessing phase. For this purpose, we use a kNN classifier. The K individual feature subsets are merged into a single feature set containing all selected features. Let $FS = (f_1, \ldots, f_S)$ be the resulting merged feature set and $op_{k,j}$ denotes the opinion of the k^{th} feature selector H_k about the selected feature f_j. This opinion is the weight assigned by H_k to feature f_j and it is equal to zero if feature f_j is not selected by H_k. A confidence level $conf_{k,j}$ is assigned to each selector H_k about each opinion $op_{k,j}$. The confidence is a weight calculated as follow:

$$conf_{k,j} = op_{k,j} * \log\left(\frac{1}{\beta_k}\right), \tag{1}$$

where β_k is the normalized error of the kNN classifier trained on the projection of the k^{th} feature subset on the data. Confidences are then normalized. Like in wrapper feature selection, the application of the classifier highlights the best performing feature subsets for that particular type of model (kNN) and have the ability to take into account feature dependencies as they consider groups of features jointly. Thus, confidence scores of all the features obtained by the same selector H_k will be affected by a common classification error score β_k corresponding to the overall subset. Thus in this step best subsets giving the minimum error rates, are also favored in addition to best individual features.

3.2.2 Reliability assessment and decision making

Given the opinions of K feature selection algorithms about the selection of a feature f_j, $Op_j = \{op_{k,j}, k = 1, \ldots, K\}$, and given the confidences associated with those opinions, $Conf_j = \{conf_{k,j}, k = 1, \ldots, K\}$, the conflict of each selector is formulated by first measuring the similarity between its opinions and those of the other selectors in the ensemble (Garcia and Puig 2003), as follows:

$$Sim_k\left(Op_j\right) = 1 - \frac{1}{(K-1)} \sum_{t=1, t \neq k}^{K} \mid op_{k,j} - op_{t,j} \mid . \tag{2}$$

Then, selector's confidence similarity with the rest of confidences, $Sim_k(Conf_j)$, is calculated the same way as in Eq. 3. Based on these calculations, the conflict raised by a selector is defined as

$$Conflict_{k,j} = Sim_k\left(Conf_j\right)\left[1 - Sim_k\left(Op_j\right)\right]. \tag{3}$$

Conflicting selectors are those with similar confidences to the agreeing selectors but completely different opinions from theirs. The conflict measure will affect selector's reliability for a feature f_j which is calculated as follows:

$$rel_{k,j} = conf_{k,j}\left(1 - Conflict_{k,j}\right). \tag{4}$$

Finally, the original opinions about the features are adjusted by multiplying them by the associated reliability factors after being normalized. The selected features are the best ranked ones according to the sum of their adjusted opinions over all selectors. The robust aggregation method is implemented using matlab software.

4 Experimental study

In this section, we compare the performance of our proposed ensemble feature selection method and those of other methods. Our experimental data consists of seven cancer diagnosis microarrays data sets.

Table 1 Datasets characteristics

Dataset	No. of samples	$\#Class1$	$\#Class2$	No. of features	Reference
DLBCL	77	58	19	7029	Shipp et al. (2002)
Bladder	31	11	20	3036	Dyrskjot et al. (2003)
Lymphoma	45	22	23	4026	Alizadeh et al. (2000)
Prostate	102	52	50	12600	Singh et al. (2002)
Breast	97	46	51	24482	van 't Veer et al. (2002)
CNS	60	21	39	7129	Pomeroy et al. (2002)
Lung	181	31	150	12533	Gordon et al. (2002)

4.1 Data sets

Seven gene expression data sets are used for experiments. They typically consist of several thousands of features and tens of instances. The classification in these data sets is binary and its task is cancer diagnosis. Table 1 summarizes the characteristics of the seven data sets. The Lymphoma data set contains missing values for numeric attributes that we replace using the kNN imputation method proposed by Troyanskaya et al. (2001). This method takes advantage of the correlation structure in the data and uses the average of records that have similar completed data patterns to impute missing values.

4.2 Performance metrics

We use tenfold stratified CV to evaluate classification and stability performances of the different ensemble feature selection methods. At each iteration of the stratified CV, feature selection is performed using the training part of the data and a classifier is then applied to evaluate the prediction performance. According to Saeys et al. (2008), for high dimensional and small sample size data, only a small amount of best features returned by the feature selector is sufficient and relevant for classification.They used only 1% in their experiments for feature selection on similar data sets, thus we also choose this pre-defined percentage of features in our experiments. As we used the classification performance of kNN classifier to guide the selection of a consensus feature set in the proposed ensemble feature selection method, we use the same classifier for the final classification performance evaluation also, as one of the goals is to optimize accuracy. Nevertheless, other classifiers could also be used for this purpose. The distance metric used for kNN is the Euclidean distance and the number of nearest neighbors for this algorithm was set to 1, after experimental evaluations of different values of k and based on the obtained results. We evaluate also the stability to compare our proposed method to other existing ensemble feature selection methods.

Classification performance The experimented data sets contain imbalanced classes (Table 1), thus a model can assign the value of the majority class for all predictions and achieve a very high classification accuracy. This model is not of interest for the

Table 2 Tenfold CV F-measure with kNN classifier and full feature sets

Dataset	DLBCL	Bladder	Lymphoma	Prostate	Breast	CNS	Lung
Fm	0.9074	0.8571	0.7442	0.7327	0.5581	0.4681	0.9677

considered problems. Consequently, alternative measures to the classification accuracy have been proposed. Among them, F-measure which is interpreted as a weighted average of the precision and recall which consider only one class (minority or majority). The precision is the percentage of positive predictions that are correct. The Recall (or sensitivity) is the percentage of positive labeled instances that were predicted as positive. The F-measure reaches its best value at 1 and worst score at 0. To make some comparisons more precise, we also use the McNemar's statistical test which is applied to test whether a proposed algorithm significantly outperforms others on a given data set. The null hypothesis is accepted when the McNemar's test is less than 3.841459 or with a p value greater than 0.05, otherwise we reject the null hypothesis in favor of the alternative hypothesis that the two algorithms have different performance.

Stability The stability of a feature selection algorithm is the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution (Kalousis et al. 2007). We measured stability on the different selected feature subsets (SFS) obtained for the tenfolds of the used CV. As said before, the top 1% best features of the rankings obtained by the different selectors, are chosen to evaluate classification but also stability. In Kuncheva (2007), authors propose a stability index to measure to which extent K sets of selected features of size s share common features:

$$Stab\,(S_1, \ldots, S_K) = \frac{2}{K\,(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left(|S_i \cap S_j| - \frac{s^2}{d} \right) / \left(S - \frac{s^2}{d} \right), \quad (5)$$

where d is the total number of features, and S_i, S_j are two different feature sets. The ratio $\frac{s^2}{d}$ corrects the bias of selecting common features in both sets by chance. This index satisfies $-1 < Stab \leq 1$ and the greater is its value the larger is the number of commonly selected features in various sets.

4.3 Results analysis

The classification performance and stability of our proposed method (RAA) are compared to several ensemble aggregation techniques discussed before. We report also the classification performance of ensemble classifier aggregation referred to as ECA, which instead of combining all the selected subsets, it aggregates decisions of classifiers built on each individual SFS. The aggregation technique used for ECA is the simple and efficient majority vote aggregation method that aggregates class labels. Note that ECA has not a corresponding stability performance, as it is built using several feature subsets and not a single one. Its unique objective is to enhance predic-

Table 3 Tenfold CV F-measure and stability of homogeneous ensembles with relief

Dataset		Relief	ECA	RAA	WMA	OFA	CLA	CAA	RRA
DLBCL	Fm	**0.9643**	0.9550	0.9455	0.9358	**0.9643**	0.8972	0.9550	0.9464
	Stab	0.6278	–	**0.8511**	0.8435	0.8416	0.4490	0.8281	0.6288
Bladder	Fm	0.6667	**0.7619**	0.5882	0.5263	0.6316	0.6316	0.7368	0.2667
	Stab	0.4335	–	0.6110	**0.6206**	0.5814	0.1296	0.5415	0.4779
Lymph	Fm	0.8980	**0.9362**	0.8571	0.8571	0.8511	0.7826	0.8750	0.9200
	Stab	0.5095	–	**0.7777**	0.7582	0.6902	0.1717	0.7381	0.4432
Prostate	Fm	0.7736	0.7475	**0.7843**	0.7723	0.7723	0.7573	**0.7843**	**0.7843**
	Stab	0.6371	–	**0.8879**	**0.8881**	0.8869	0.6079	0.8822	0.6779
Breast	Fm	0.4000	0.4750	0.5176	0.4819	0.4762	**0.5238**	0.4762	0.4524
	Stab	0.3363	–	0.6545	0.6229	0.6624	0.1682	**0.6652**	0.4859
CNS	Fm	0.4167	0.4898	**0.5185**	0.4490	0.4528	0.3333	0.5000	0.4706
	Stab	0.4953	–	**0.8402**	**0.8458**	0.8033	0.4169	0.8020	0.6386
Lung	Fm	0.9524	**0.9688**	0.9524	**0.9688**	**0.9688**	0.9206	0.9524	0.9375
	Stab	0.7982	–	**0.9003**	**0.9024**	0.8739	0.6750	0.8822	0.6934

Predictive performance and Kuncheva index for stability measure are evaluated on the subset of 1% top ranked features and best results are highlighted in bold face

tive performance. For the heterogeneous ensembles with data variation (Table 7), we also compare proposed approaches to the Random Forest (RF), introduced by Breiman (2001), as it is an embedded method that also employ both, construction of sub-models for different random feature subsets as well as data perturbation via bootstrapping. In RF, feature importance is calculated by permuting each feature and measuring how much the permutation decreases the accuracy of the model. Given the high dimensionality of the data in our feature selection experiments, some features can be missed if we use a small number of trees. Thus, we set the forest size to 1000 trees in order to have a sufficiently large number of decision trees to select the most relevant features for good classification estimates. At each candidate split in the RF learning process, we set the size of the random subset of feature to $\sqrt{d}$, the typical number used for a classification problem with d features according to Hastie et al. (2009). Tables 3, 4, 5, 6 and 7 show the F-measure (Fm) and stability (Stab) results obtained from all settings for the seven data sets. Note that reported stability results concern the final SFS used for the final classification. For all experimented methods, the Fm and stability values reported are obtained using 1% of the original set of features.

To have an overview on the results, we first report in Table 2 the tenfold CV classification performance (Fm) of kNN algorithm on the seven data sets using the full feature sets and without applying a feature selection.

4.3.1 Homogeneous ensembles

Homogeneous ensembles with Relief

Table 3 shows that ensemble methods improve the baseline classification performance for most cases except for Bladder, where all Fm values are worse than that obtained

Table 4 Tenfold CV F-measure and stability of homogeneous ensembles with mRMR

Dataset		mRMR	ECA	RAA	WMA	OFA	CLA	CAA	RRA
DLBCL	Fm	0.9369	0.9558	**0.9643**	**0.9643**	0.9558	0.8718	**0.9643**	0.9474
	Stab	0.5726	–	0.5881	0.5991	**0.6112**	0.0329	0.5859	0.1692
Bladder	Fm	0.8800	**0.9167**	**0.9167**	**0.9167**	0.7000	**0.9167**	0.5714	0.8696
	Stab	0.4742	–	0.5037	0.5015	0.0527	0.5015	0.1696	**0.5244**
Lymph	Fm	**0.9583**	0.9362	0.9167	0.9167	0.8980	0.6818	0.9388	0.8936
	Stab	0.6353	–	**0.7043**	0.6796	**0.7054**	0.0315	0.7082	0.2391
Prostate	Fm	0.8454	**0.8866**	0.8776	0.8687	**0.8889**	0.7800	0.8660	0.8632
	Stab	**0.7203**	–	0.6971	0.6967	0.7083	0.0483	0.6957	0.2655
Breast	Fm	**0.7033**	0.6742	0.6947	0.6739	0.6737	0.6667	0.6875	0.6170
	Stab	**0.4975**	–	0.4362	0.4436	0.4559	0.0234	0.4350	0.1013
CNS	Fm	0.4681	0.6000	0.5455	0.5532	0.4186	**0.6500**	0.5116	0.4186
	Stab	**0.4117**	–	0.3510	0.3469	0.3706	0.0159	0.3292	0.0683
Lung	Fm	**0.9677**	**0.9677**	0.9153	0.9153	0.9333	0.9153	0.9153	**0.9677**
	Stab	**0.8314**	–	0.8214	0.8220	0.8172	0.0973	0.8242	0.3600

Predictive performance and Kuncheva index for stability measure are evaluated on the subset of 1% top ranked features and best results are highlighted in bold face

with the full feature set, and Breast cancer data set. ECA gives the best Fm for three data sets and this improvement is significant for Lymphoma data set, with a McNemar's test equal to 6.05 (larger than 3.841459) and a p value of 0.01 which means that the null hypothesis is rejected at 5% significance level for this data set. However, the improvement of ECA over the baseline is not significant for Bladder data set with a small McNemar's test equal to 0.5 and a p value of 0.4795. RAA then WMA are performing well by giving similar Fm results as the baseline learner but with much better stability of feature selection. In terms of stability of feature selection, results show that Relief clearly benefits from the ensemble version, especially with RAA and WMA which give the best stability results for most data sets. CAA and OFA give also good results. It is noticed that CLA gives poor stability results. This technique relies on feature ranking. Therefore, RAA and most selector ensembles are efficient with this setting especially in terms of stability of feature selection.

Homogeneous ensembles with mRMR

Table 4 shows that ECA achieves the best classification results for three data sets. For this setting also, RAA and WMA have a similar behaviour, they perform well by giving classification results superior to mRMR for four data sets, but this is not supported by statistical test for DLBCL data set with p values equal to 1.00, 0.24 and 0.24 respectively for ECA, RAA and WMA and a same p value equal to 0.47 for Bladder data set. These ensemble methods conserve approximately the baseline stability results. Performances of other selector ensembles vary depending on the data set and do not improve the baseline results.

Table 5 Tenfold CV F-measure and stability of homogeneous ensembles with *t* test

Dataset		*t* test	ECA	RAA	WMA	OFA	CLA	CAA	RRA
DLBCL	Fm	0.9174	0.9273	0.9074	0.9074	0.9009	0.9217	0.8909	**0.9381**
	Stab	**0.8679**	–	0.7490	0.4585	0.8340	0.0987	0.8280	0.3867
Bladder	Fm	0.8696	**0.9565**	0.8696	0.8182	0.8333	0.8000	0.8333	0.9091
	Stab	**0.8098**	–	0.7211	0.6790	0.6915	0.0616	0.6819	0.2450
Lymph	Fm	0.9362	0.9362	0.9167	**0.9565**	0.9362	0.7917	0.9362	**0.9565**
	Stab	**0.8008**	–	0.6925	0.4406	0.7239	0.0422	0.7419	0.3283
Prostate	Fm	**0.8776**	**0.8776**	0.8687	0.8687	0.8687	0.6809	0.8571	0.8660
	Stab	**0.8050**	–	0.7440	0.7281	0.7556	0.1131	0.7514	0.3840
Breast	Fm	0.6434	0.6434	0.6434	0.6434	0.6434	0.6434	0.6434	0.6434
	Stab	1	–	1	1	1	1	1	1
CNS	Fm	0.4167	**0.5366**	0.4286	0.5116	0.4348	0.4651	0.4348	0.4286
	Stab	**0.4834**	–	0.3020	0.1708	0.3807	0.0329	0.3247	0.1037
Lung	Fm	0.9063	0.8923	0.9063	0.9206	0.9063	**0.9333**	0.9063	0.9231
	Stab	**0.8953**	–	0.8470	0.8342	0.8607	0.2260	0.8589	0.5109

Predictive performance and Kuncheva index for stability measure are evaluated on the subset of 1% top ranked features and best results are highlighted in bold face

Homogeneous ensembles with *t* test
The results reported in Table 5 show that *t* test filter benefits from some ensemble version to improve classification performance. ECA, RRA are the most efficient ensemble versions for this purpose and RAA conserves very similar results. However, this classification improvement is not statistically significant, and it is coupled with stability decrease compared to the baseline stability.

4.3.2 Heterogeneous ensembles

Table 6 shows the performance results of the heterogeneous ensembles trained on the same data. It can be observed that Relief is the baseline algorithm that benefits the most from the heterogeneous ensemble version especially comparing to Fm and stability results of OFA. The heterogeneous ensemble method using OFA improves also stability results of mRMR but with a classification performance sacrifice in some cases. Even if RAA , WMA or RRA sometimes improve the baseline classification performance (Bladder, Lymphoma and Lung data sets), *t* test remains the algorithm that achieves the best classification-stability trade-off in this setting. Stability performance of ensemble methods are smaller than *t* test baseline method for all data sets and OFA gives the best stability results for the ensemble methods.

For the heterogeneous ensembles with data variation, results reported in Table 7 among all algorithms, *t* test baseline gives the best apparent trade-off classification performance-stability for five data sets. For Lung data set, RAA and WMA or *t* test can be chosen depending on the preferred evaluation criterion. If we look at classification performances only, we notice that the best Fm measures are obtained either by mRMR,

Table 6 Tenfold CV F-measure and stability of heterogeneous ensembles

Dataset		Relief	mRMR	*t* test	ECA	RAA	WMA	OFA	CLA	CAA	RRA
DLBCL	Fm	0.9381	**0.9739**	0.9107	0.9558	0.9483	0.9391	0.9204	0.9286	0.9369	0.9464
	Stab	0.6949	0.5840	**0.8543**	–	0.5928	0.5517	0.7329	0.5666	0.6298	0.6077
Bladder	Fm	0.6667	0.8696	0.8182	0.9091	**0.9167**	**0.9167**	0.8696	**0.9167**	0.6400	**0.9167**
	Stab	0.5954	0.4550	**0.8128**	–	0.3906	0.3914	0.6331	0.4912	0.5829	0.5045
Lymph	Fm	0.8511	0.9167	0.9362	0.9362	**0.9583**	0.9388	0.8980	0.9167	0.9200	**0.9583**
	Stab	0.5421	0.5999	**0.7492**	–	0.5921	0.5881	0.5915	0.4574	0.5253	0.5185
Prostate	Fm	0.7525	0.8632	**0.8776**	0.8542	0.8200	0.8350	0.8317	0.8039	0.8283	0.8235
	Stab	0.6854	0.7116	**0.7793**	–	0.6587	0.6583	0.7139	0.6262	0.5708	0.6831
Breast	Fm	0.4384	0.6522	0.6434	**0.6667**	0.6434	0.3896	0.5060	0.6098	0.6364	0.6098
	Stab	0.5253	0.4801	**1.0000**	–	**1.0000**	0.5253	0.6524	0.5754	0.2565	0.5752
CNS	Fm	0.4898	**0.5417**	0.3478	0.4681	0.4000	0.3902	0.5217	0.5000	0.4545	0.4783
	Stab	**0.5100**	0.3608	0.4736	–	0.3181	0.3269	0.4335	0.3459	0.2258	0.3706
Lung	Fm	0.9538	0.9677	0.9231	0.9688	0.9688	**0.9841**	**0.9841**	0.9180	0.9688	0.9524
	Stab	0.7885	0.7876	**0.8784**	–	0.6112	0.5757	0.7894	0.7411	0.7732	0.7849

Predictive performance and Kuncheva index for stability measure are evaluated on the subset of 1% top ranked features and best results are highlighted in bold face

ECA or RF algorithm. However, the latter gives often poor feature selection stability. We notice that stability results of heterogeneous ensembles with data variation are not better than those obtained with same data in spite of the highest ensemble size, equal to 30 with varying the data, against an ensemble size of 3 with varying only the base learner. This proves that here, stability is mainly affected by the algorithm variation. However, the data variation in this setting affected the classification results by a general performance improvement over the ensemble methods.

4.4 Discussion

For all experiments, ECA gives often good classification performances. However, this is not good enough, since there is not a corresponding stability performance. In fact, the objective of ECA is not to have a stable feature selection but to enhance predictive performance by aggregating classifier results built on different feature subsets. Therefore, if the interest is in classification performance, ECA is the technique to use to get good results. However, if we search for techniques to achieve good classification and feature selection stability at the same time, RAA, our proposed method based on conflict resolution and reliability assessment, is a good solution if it is applied with homogeneous ensembles formed with instable baseline learners. Relief is such an algorithm which benefits a lot from the ensemble version. WMA, which is a simple technique that aggregates feature weights, has proved its efficiency with the same settings. OFA is also an ensemble feature selection method that achieves a well trade-off classification performance-stability in many settings. The results of the ensemble

Table 7 Tenfold CV F-measure and stability of heterogeneous ensembles with data variation

Dataset		RF	Relief	mRMR	*t* test	ECA	RAA	WMA	OFA	CLA	CAA	RRA
DLBCL	Fm	0.9421	0.9358	**0.9649**	0.9107	0.9464	0.9369	0.9369	0.9464	0.9009	0.9550	0.9027
	Stab	0.4524	0.6301	0.5928	**0.8682**	–	0.6807	0.6434	0.7212	0.1667	0.6927	0.6086
Bladder	Fm	0.9091	0.4444	**0.9167**	0.9091	**0.9167**	0.6087	0.5263	0.8571	0.8000	0.8333	0.7368
	Stab	0.6243	0.4224	0.4539	**0.8369**	–	0.5579	0.5608	0.4696	0.0768	0.4247	0.4262
Lymph	Fm	0.9565	0.8511	**0.9778**	0.9565	0.9388	0.7917	0.7826	0.9362	0.8511	0.9362	0.8980
	Stab	0.5180	0.4445	0.6515	**0.7952**	–	0.1566	0.1314	0.6712	0.0601	0.6212	0.2902
Prostate	Fm	**0.9109**	0.7677	0.8980	0.8889	0.8687	0.7677	0.7677	0.8119	0.7723	0.7451	0.7379
	Stab	0.3854	0.6288	0.7277	0.8020	–	**0.8194**	0.7948	0.6773	0.1772	0.6769	0.5776
Breast	Fm	**0.6593**	0.4419	**0.6593**	0.6434	0.6239	0.6024	0.5542	0.5102	0.4000	0.6465	0.5806
	Stab	0.3518	0.4237	0.4903	**1.0000**	–	0.2065	0.3820	0.9838	0.3639	0.7216	0.9925
CNS	Fm	0.1429	0.3158	0.4889	0.5000	**0.5789**	0.5385	0.4490	0.4082	0.3721	0.4444	0.3111
	Stab	0.4705	0.4683	0.4335	0.5005	–	0.5062	0.3959	0.5429	0.0234	0.4825	0.1278
Lung	Fm	**0.9836**	0.9524	0.9677	0.8955	**0.9841**	0.9538	0.9538	0.9688	0.9355	0.9688	0.9180
	Stab	0.3112	0.7864	0.8282	**0.9027**	–	0.8539	0.8611	0.7980	0.2963	0.7774	0.8102

Predictive performance and Kuncheva index for stability measure are evaluated on the subset of 1% top ranked features and best results are highlighted in bold face methods are sensitive to the applied base learners, and are not efficient to improve performances of stable algorithms such as *t* test.

5 Conclusion

In this paper, we proposed an ensemble feature selection approach based on feature selectors' reliability assessment. The interest of this approach is that it aims at providing a unique and stable feature selection without ignoring the predictive accuracy aspect. First, different subsets of features are obtained by homogeneous ensembles or heterogeneous ensembles. Then, we proposed a robust aggregation technique based on classification performance and reliability assessment to combine selectors' ensemble output. We compared our proposed approach to several existing techniques and to individual feature selection results. Experiments showed that our approach often improves classification performance and stability for high dimensional and small sample size data sets or at least maintains the baseline results when they are specially high. To enhance stability, the homogeneous ensembles formed with instable base learners are better than heterogeneous ensembles as they yield optimal stability results. The comparative study on ensemble feature selection methods and the proposed robust aggregation technique could be extended to other feature selection methods. Studying the relationship between the baseline algorithm used for the creation of the selector ensemble and the ensemble aggregation mechanism would be interesting to further improve stability of ensemble feature selection.

References

Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3):392–398

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson JJ, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769):503–511

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chan D, Bridges SM, Burgess SC (2008) An ensemble method for identifying robust features for biomarker discovery. Chapman and Hall/CRC Press, Boca Raton

Dietterich TG (2000) Ensemble methods in machine learning. In: Proceedings of the first international workshop on multiple classifier systems. Springer-Verlag, London, UK, UK, pp 1–15

Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft TF (2003) Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet. 33:90–96

Garcia MA, Puig D (2003) Robust aggregation of expert opinions based on conflict analysis and resolution. In: CAEPIA, Lecture Notes in Computer Science, Springer, pp 488–497

Gordon G, Jensen R, Hsiao L, Gullans S, Blumenstock J, Ramaswamy S, Richards W, Sugarbaker D, Bueno R (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 62:4963–4967

Gosset WS (1908) The probable error of a mean. Biometrika 1:1–25

Guyon I, Elisseff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hastie TJ, Tibshirani RJ, Friedman JH (2009) The elements of statistical learning : data mining, inference, and prediction. Springer series in statistics. Springer, New York

Kalousis A, Prados J, Hilario M (2007) Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. 12(1):95–116

Kira K, Rendell L (1992) A practical approach to feature selection. In: Sleeman D, Edwards P (eds) International conference on machine learning, pp 368–377

Kohane IS, Kho AT, Butte AJ (2003) Microarrays for an integrative genomics. MIT Press, Cambridge

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol 2, Morgan Kaufmann Publishers Inc., pp 1137–1143

Kolde R, Laur S, Adler P, Vilo J (2012) Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 28(4):573–580

Kuncheva L (2007) A stability index for feature selection. In: Proceedings of the 25th IASTED international multi-conference: artificial intelligence and applications, Innsbruck, Austria, pp 390–395

Mitchell L, Sloan T, Mewissen M, Ghazal P, Forster T, Piotrowski M, Trew A (2014) Parallel classification and feature selection in microarray data using sprint. Concurr Comput Pract Exp 26(4):854–865

Okun O (2011) Feature selection and ensemble methods for bioinformatics: algorithmic classification and implementations. IGI Global, Hershy, PA

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415(6870):436–442

Saeys Y, Abeel T, Peer Y (2008) Robust feature selection using ensemble feature selection techniques. In: Proceedings of the European conference on machine learning and knowledge discovery in databases—Part II, ECML PKDD '08, Springer-Verlag, Berlin, Heidelberg, pp 313–325

Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507–2517

Schowe B, Morik K (2011) Fast-ensembles of minimum redundancy feature selection. In: Ensembles in machine learning applications: studies in computational intelligence, vol 373, pp 75–95

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS (2002) Diffuse large b(cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 9:68–74

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1(2):203–209

Troyanskaya OG, Cantor M, Sherlock G, Brown PO, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for dna microarrays. Bioinformatics 17(6):520–525

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002, January) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536