Improving Stability of Feature Selection for Brain Tumour Diagnosis Using ^{1}H-MRS Data

Albert Vilamala and Lluís A. Belanche

Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3. 08034, Barcelona, Spain
{avilamala,belanche}@lsi.upc.edu

Abstract. Magnetic Resonance Spectroscopy for brain tumour diagnosis is progressively replacing harmful biopsy. Nonetheless, dealing with such multidimensional outcome becomes a difficult task for the medical community. Computation-based tools able to effectively reduce dimensionality of data without losing diagnostic ability ease the interpretation of results. The current study presents a novel technique to improve stability of feature subset selection algorithms by means of an instance weighting approach. We report experiments performed on real data showing an improvement on feature selection stability up to 40%.

Keywords: Feature Selection, Stability, Instance Weighting, Magnetic Resonance Spectroscopy

1 Introduction

Diagnosis of brain tumours from Magnetic Resonance Spectroscopy (MRS) is a non-invasive technique aiming at substituting the gold-standard but unpleasant biopsy. The combination of MRS results with Machine Learning (ML) solutions is a promising tool to accurately diagnose new patients. However, practitioners often face the problem of building reliable models using only few available instances (records of patients), which moreover are made up of a large set of features. The difficulty of interpreting models with large numbers of features is alleviated by the use of Feature Selection (FS) techniques, which aim at picking the very relevant features explaining the target concept of interest. Nevertheless, the confidence in a model highly depends on its *stability* with respect to changes in the data used to obtain it, either in the instances themselves or in the features used to describe them. For example, in a typical FS process using cross-validation, different features are typically selected in every validation fold.

Previous research interest in this direction has focused mainly in assessing the performance of different FS algorithms in terms of feature subset stability, leading to the development of measures to properly evaluate it [1, 2]; few works address the explicit improvement of such stability, notably an Ensemble-base FS [3] and a Group-base FS [4]. More recently, Han et al. [5] coupled the hypothesis margin with an importance sampling approach to diminish the small sample size problem resulting in more stable subsets of features.

In the current study we present a novel method that aims at providing a more stable selection of feature subsets when variations in the training process occur. This is accomplished by *weighting* the instances according to their outlying behavior; this weighting is a preprocessing step that is independent of the learner or the specific FS algorithm. We report performance in two series of experiments: first using microarray gene expression datasets and then brain tumour MRS data for the diagnosis of two severe pathologies (glioblastomas and metastases) [6, 7]. Our results show increases in FS stability up to a 40%.

2 A new Instance Weighting method

Let $D = \{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$ be a training data set of length N, each instance $\boldsymbol{x}_i \in \mathbb{R}^d$ with its corresponding class label t_i. The proposed method assesses the importance of each instance according to its outlying behavior before applying any FS strategy. In particular, the approach is based on the *hypothesis margin* concept, which states that *"the margin of an hypothesis with respect to an instance is the distance between the hypothesis and the closest hypothesis that assigns alternative label to the given instance"* [8]. The margin of a hypothesis $\mathbf{x} \in \mathbb{R}^d$ can be calculated as

$$\theta(\mathbf{x}) = \frac{1}{2}\left(\|\mathbf{x} - m(\mathbf{x})\| - \|\mathbf{x} - h(\mathbf{x})\|\right). \tag{1}$$

A single outlier in a neighborhood might mislead the margin calculus of all its neighbors. With the purpose of obtaining a more robust evaluation, the average margin between every instance in D and all the rest can be calculated:

$$\theta(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M}\|\mathbf{x} - m_i(\mathbf{x})\| - \frac{1}{H}\sum_{i=1}^{H}\|\mathbf{x} - h_i(\mathbf{x})\|, \tag{2}$$

being $m_i(\mathbf{x})$ and $h_i(\mathbf{x})$ the i-th nearest *miss* (instance of different class) and i-th nearest hit (instance of same class) in D, respectively; where M, H are the total number of misses and hits (such that $M + H + 1 = N$).

Instances $\mathbf{x}$ achieving highly positive $\theta(\mathbf{x})$ present good modeling behavior (being far from misses and close to hits), while those with highly negative $\theta(\mathbf{x})$ become outlying ones (surrounded by misses and far from hits). The presence or absence of these latter instances is therefore a source of unstability. In order to obtain a bounded positive weight in $(0, 1)$, we use a logistic function:

$$\omega(\mathbf{x}) = \frac{1}{1 + \exp\{-\alpha\, z\,(\theta\,(\mathbf{x}))\}}, \tag{3}$$

where α is a parameter controlling the slope, and $z(\cdot)$ is the *standard score* $z(x) = (x - \hat{\mu}_D)/\hat{\sigma}_D$, being $\hat{\mu}_D$ and $\hat{\sigma}_D$ the sample mean and standard deviation of $\theta(\mathbf{x})$, for all $\mathbf{x} \in D$, respectively. A suitable value for α will depend on the

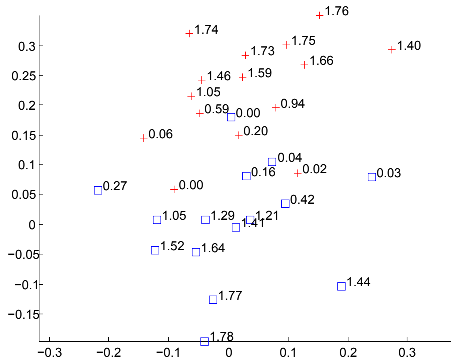


Fig. 1: Ratings assigned by RLIW to the synthetic illustrative instances.

problem and the user's needs. As a default value, we propose to set $\alpha = 3.03$, which corresponds to assign a weight of 0.954 to an instance whose average margin is two standard deviations from the mean, that is $\theta(\mathbf{x}) = 2\hat{\sigma}_D$.

The proposed *Reweighted Logistic Instance Weighting* (RLIW) performs feature selection by repeatedly applying Eqs. (2), (3) to compute the ω weights, use them in a weighted FS algorithm, removing the worst feature (or features), re-compute the ω weights, etc, until a stopping criterion is met.

An example. In order to illustrate the RLIW procedure, a simple example is provided. Let $\mathbf{X} \in \mathbb{R}^{N \times F}$ be a synthetic dataset, where $N = 30$ is the number of instances, and $F = 2$ the number of features; and $\mathbf{y} \in \{1, -1\}^N$ is the vector of labels providing the class membership of every instance. Each instance was obtained by equally sampling from one of two distributions: either $\mathbf{x} \sim \mathcal{N}(\mu_1, \Sigma)$ or $\mathbf{x} \sim \mathcal{N}(\mu_2, \Sigma)$, where $\mu_1 = [0, 0], \mu_2 = [0, 0.25]$ and $\Sigma = \left[\begin{smallmatrix} 0.01 & 0.00 \\ 0.00 & 0.01 \end{smallmatrix}\right]$; and labeled according to the distribution it comes from.

Fig. 1 shows the dataset with the weighting obtained by RLIW, which clearly assigns low values to instances close to the boundary between classes and those inside opposite-class region; and assigns higher values the farther from the boundary inside the proper-class region. This is consistent with the intuition that outlying instances are a source of unstability and therefore, must be lowly rated.

2.1 Embedding into a Feature Selection algorithm

The weights in Eq. (3) need to be supplied to FS algorithms capable to accept them to improve its selection ability. Here we present two different approaches.

SVM-RFE. SVM-RFE (Support Vector Machine – Recursive Feature Elimination) [9] is a backward selection algorithm that uses the weights of a linear SVM to rank the remaining features at each iteration. It starts up by using all the available features to train a soft-margin SVM optimizing the objective function

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i,$$

where ξ is a vector of slack variables or deviations from the hyperplane, C is the hyperparameter that controls the trade-off between separating with maximal margin and allowing missclassifications; $\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$ is the weight vector, being $\alpha, \mathbf{y}, \mathbf{x}$ the vectors of Lagrangian parameters, class labels and instances, respectively. Once convergence is achieved, the weight vector is used to compute the ranking criterion for each feature as $c_j = |w_j|$. Those features with the lowest rank are discarded, starting the next iteration using only the remaining ones.

In this study we use an extension of the SVM that inserts instance weights to multiply the slack variable in the objective function [5]:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \omega_i \xi_i,$$

where $\omega_i = \omega(\mathbf{x}_i)$ is the weight assigned to the i-th instance, according to Eq. 3.

RelievedF-RFE. Relief is a family of FS filters that use the hypothesis-margin concept in Eq. 1 to assess the importance of each feature in a dataset D as the accumulated influence that each feature has in computing the margin of every instance in D. In particular, RelievedF [10] is a deterministic feature ranking algorithm that depends on a user-defined parameter k. The algorithm picks one instance at a time and computes the hypothesis margin of each feature independently, accumulating the feature-wise distances to the k nearest hits and k nearest misses. As a result, the weight $W(j)$ given to feature j is its average distance to the selected neighbors:

$$W(j) = \sum_{i=1}^{N} \frac{1}{k} \sum_{l=1}^{k} \left(|\mathbf{x}_{i,j} - m_l(\mathbf{x}_i)_j| - |\mathbf{x}_{i,j} - h_l(\mathbf{x}_i)_j|\right).$$

Using this weighting strategy, features can be ranked and those ones with lowest rank can be removed. Feature subset selection can be obtained by repeatedly applying the previous equation while removing the worst feature (or features) at a time, obtaining RelievedF-RFE. In order to be able to use instance weights, we make use of a weighted version [5] of RelievedF:

$$W(j) = \sum_{i=1}^{N} \omega_i \sum_{l=1}^{k} \left(\omega_{i,l}^{M} |\mathbf{x}_{i,j} - m_l(\mathbf{x}_i)_j| - \omega_{i,l}^{H} |\mathbf{x}_{i,j} - h_l(\mathbf{x}_i)_j|\right), \tag{4}$$

where $\omega_i = \omega(\mathbf{x}_i)$, $\omega_{i,l}^{M} = \omega(m_l(\mathbf{x}_i))$ and $\omega_{i,l}^{H} = \omega(h_l(\mathbf{x}_i))$, obtained in Eq. 3. Using this *double* weighting strategy, features can be ranked according to their importance, and at the same time favoring stability due to the ω weights.

2.2 Comparison with previous approaches

The study proposed in [5] not only provided a theoretical analysis on the bias-variance decomposition of the FS error –showing the suitability of Instance Weighting (IW) for reducing variance– but also supplied an empirical framework to effectively calculate the importance of each instance. Specifically, the strategy consists in two basic steps: first, all instances are mapped into a new *margin vector feature space* (MVFS) where each coordinate is calculated according to the equation:

$$\mathbf{x}'_{i,j} = \sum_{l=1}^{M} |\mathbf{x}_{i,j} - m_l(\mathbf{x}_i)_j| - \sum_{l=1}^{H} |\mathbf{x}_{i,j} - h_l(\mathbf{x}_i)_j|. \tag{5}$$

Then, using this new coordinate system, the importance of each instance is calculated:

$$\omega(\mathbf{x}) = \frac{1/\bar{d}(\mathbf{x}')}{\sum_{i=1}^{N} 1/\bar{d}(\mathbf{x}'_i)}, \tag{6}$$

where

$$\bar{d}(\mathbf{x}') = \frac{1}{N-1} \sum_{p=1, \mathbf{x}'_p \neq \mathbf{x}'}^{N-1} \|\mathbf{x}' - \mathbf{x}'_p\|. \tag{7}$$

A first concern with this proposal arises when mapping to the new MVFS. According to Eq. 5, a new coordinate is derived for each dimension of an instance; given that in Eq. 7 all dimensions are considered at a time by means of Euclidean distance, there is no need for an explicit calculation of each new coordinate.

Second, the imposition of a normalization factor in Eq. 6 –such that the sum of all weights adds to 1– might lead to undesirable effects. In a hypothetical setting where N instances are equally separated at distance 1 in the MVFS, the assigned weight should be innocuous and directly comparable to standard FS without IW (Std-FS). However, due to the normalization, every instance is weighed as $1/N$. The undesirable side effect can be clearly appreciated in the case of SVM-RFE, where the value of the C parameter is reduced by a factor of N. Furthermore, the relative difference in weighting between two instances is minimized, negatively affecting weighted FS techniques, such as RelievedF-RFE.

3 Experimental Work

3.1 Datasets used

Han & Yu Synthetic Data. A synthetic dataset previously used to verify the performance of stable feature subset strategies [5] is also employed in this study. It consists of $M = 500$ training sets, each of the form $\mathbf{X}^m \in \mathbb{R}^{N \times F}$, with $N = 100$ instances and $F = 1,000$ features, for $m = 1, \ldots, M$. Every instance is equiprobably drawn from one of two distributions: $\mathbf{x} \sim \mathcal{N}(\mu_1, \Sigma)$ or $\mathbf{x} \sim \mathcal{N}(\mu_2, \Sigma)$, where

$$\mu_1 = (\underbrace{0.5, ..., 0.5}_{50}, \underbrace{0, ..., 0}_{950}), \quad \mu_2 = -\mu_1,$$

and

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{100} \end{bmatrix},$$

being $\Sigma_i \in \mathbb{R}^{10\times10}$, with 1 in its diagonal elements and 0.8 elsewhere. Class labels are assigned according to the expression:

$$\mathbf{y}_i = \operatorname{sgn}\left(\sum_{j=1}^{F} \mathbf{X}_{i,j}\mathbf{r}_j\right), \quad \mathbf{r} = (\underbrace{0.02, ..., 0.02}_{50}, \underbrace{0, ..., 0}_{950}).$$

Real Microarray Data. A widely-used collection of microarray datasets presenting a variety of diseases will be of use. In particular, Colon [11] and Leukemia [12] cancer datasets are directly employed ($N = 62$, $F = 2000$; $N = 72$, $F = 7129$, respectively). In the case of Prostate [13] cancer dataset, a preprocessing similar as the one in [14] is performed: it consists in fixing a valid range of values for each feature to lay between $[10, 16000]$. Any value out of this interval is set to its closest limit. Afterwards, features presenting low variability ($max/min < 5$ or $max - min < 50$) are removed ($N = 102$, $F = 6034$). For the Lung [15] ($N = 181$), Breast [16] ($N = 97$) and Melanoma [17] ($N = 70$) cancer datasets, as well as for Parkinson [18] ($N = 105$) dataset, a standard t-test, keeping the 5000 top features ($F = 5000$), is applied [5].

Real MRS Data. Different datasets containing Single-Voxel Proton Magnetic Resonance Spectroscopy (SV-^{1}H-MRS) data of human brain tumours are also used to validate our method, coming from the international, multi-centre INTERPRET European project database [6]. Specifically, the two datasets consist in 78 glioblastomas (GL) and 31 metastases (ME), $N = 109$, at Long and Short Time of Echo (LTE and STE, respectively), each described by $F = 195$ features. The extreme difficulty of discriminating among these two types of tumours has been largely reported in several previous studies (see *e.g.* [19]). An independent *test set* composed of 10 GL and 30 ME (both at STE and LTE) is also used to properly validate the proposed RLIW method [7].

3.2 Figures of merit

Feature subset stability. The stability of a FS algorithm in selecting a subset of k features out of the initial sample feature size F over a batch of M runs can be evaluated using the *Kuncheva index* (KI) [1], defined as

$$\mathcal{I}_S\left(\mathcal{E}(k)\right) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \mathcal{I}_C\left(S_i(k), S_j(k)\right)$$

$$\mathcal{I}_C\left(S_i(k), S_j(k)\right) = \frac{|S_i(k) \cap S_j(k)| - (k^2/F)}{k - (k^2/F)}$$

where $S_i(k)$ is the subset of selected features of length k in the i-th run; and $\mathcal{E} = \{S_1, S_2, ..., S_M\}$ is the set containing all the retrieved feature subsets. KI values are bounded between -1 and 1, being this last one the maximum stability.

Prediction accuracy. The measure of choice for the correctness of the predictions provided by the classifiers is the average of the *balanced accuracy* or BAC [20], well suited to deal with unbalanced datasets:

$$\mathrm{BAC}\left(\mathbf{Z}\right) = \frac{1}{2M} \sum_{i=1}^{M} \left(\frac{|j \,/\, Z_{ij} = 0 \wedge Y_j = 0|}{N_0} + \frac{|j \,/\, Z_{ij} = 1 \wedge Y_j = 1|}{N_1} \right)$$

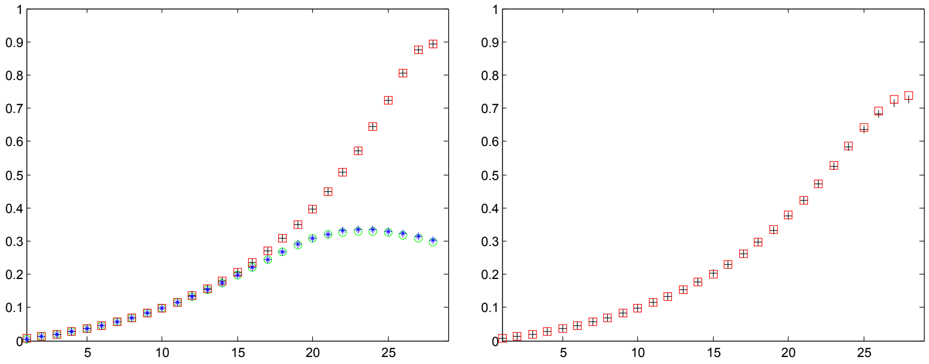
where $\mathbf{Z}_i$ are the predictions at the i-th run; $\mathbf{Y}$ is the vector of true class labels; and N_0 and N_1 are the number of samples in classes 0 and 1, respectively.

3.3 Results and Discussion

MVIW using Han & Yu synthetic data. Following [5], the experimental setting for evaluating MVIW feature subset stability consists in using each training set separately to pick the best subset of features of certain size. In particular, given a normalized multivariate training set, the importance of every instance is calculated according to Eq. 6. Then, this information is provided to a FS strategy (Section 2.1) within an RFE process, while removing the worst 10% of features per iteration until all of them have been eliminated. The procedure is repeated for each training set, calculating the KI per feature subset size.

Fig. 2a shows the feature subset stability when SVM-RFE is applied. The major increase in KI produced by MVIW (red square) compared to the Std-FS (green circle), can also be achieved by setting the parameter $C = 1/N$ without any use of IW (black cross). Contrarily, if we apply MVIW with corrected scale (blue asterisk) in order to assess the relative rating of instances, no significant improvement in stability is appreciated.

Fig. 2b shows RelievedF-RFE with parameter $k = 10$ [5]. It can be observed that incorporating MVIW (red square) presents virtually no gain in terms of feature subset stability compared to regular FS –without instance weights (black cross). For this setting, the correction of the scaling factor has not been plotted, since it does not affect the result (Eq. 4). In light of these results, we conclude that the improvement in subset stability is not caused by MVIW, but by the scaling factor applied to parameter C as a side effect.

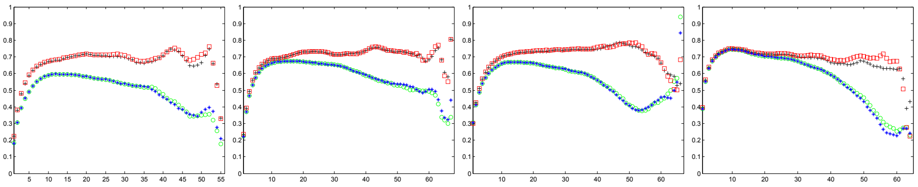


(a) SVM-RFE

(b) RelievedF-RFE

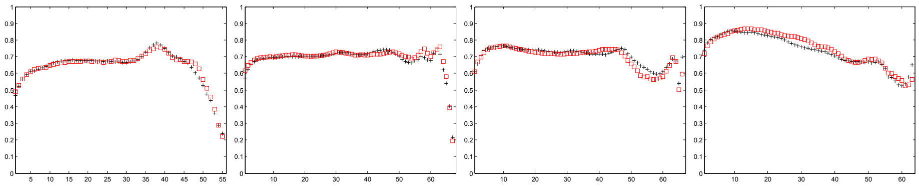
Fig. 2: Feature subset stability on Han & Yu synthetic data. The plots show the KI (vertical axis) over a set of RFE iterations (horizontal axis). For SVM-RFE, green circle corresponds to parameters $C = 1$ and $IW = none$; blue asterisk: $C = 1$ and $IW = N \times MVIW$; black cross: $C = 1/N$ and $IW = none$; red square: $C = 1$ and $IW = MVIW$. In the case of RelievedF-RFE, black cross corresponds to parameter $IW = none$; red square: $IW = MVIW$.

MVIW using microarray data. The previous effect has been verified in a larger cohort of data by designing a set of experiments on real microarray data. The same experimental settings as those stated in the previous section hold, with the difference that every dataset is analyzed by a stratified 10-times 10-fold cross-validation (10x10cv) resampling strategy, normalizing the data at every fold, in order to generate training set variability. The KI is computed per feature subset length at every inner 10cv and then computing the average over the 10 times.



(a) Colon (b) Leukemia (c) Prostate (d) Lung

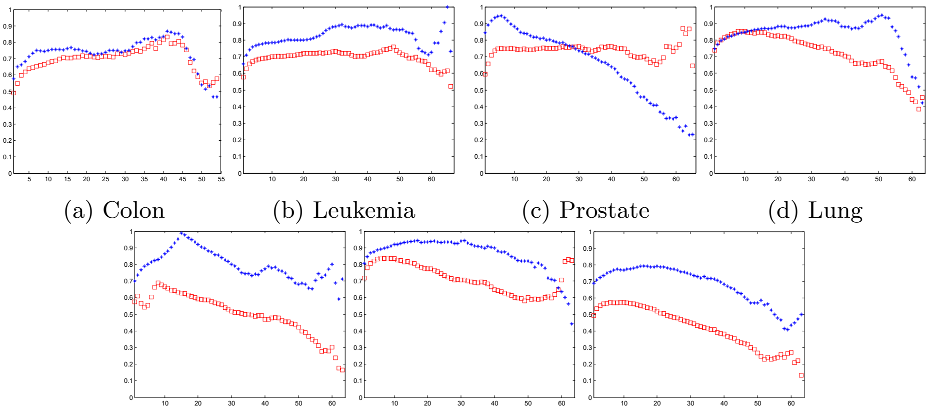
Fig. 3: Feature subset stability using SVM-RFE on real microarray data. Each plot shows the KI (vertical axis) over a set of RFE iterations (horizontal axis). Green circle corresponds to parameters $C = 1$ and $IW = none$; blue asterisk: $C = 1$ and $IW = N \times MVIW$; black cross: $C = 1/N$ and $IW = none$; red square: $C = 1$ and $IW = MVIW$.



(a) Colon (b) Leukemia (c) Prostate (d) Lung

Fig. 4: Feature subset stability using RelievedF-RFE on real microarray data. Each plot shows the KI (vertical axis) over a set of RFE iterations (horizontal axis). Black cross: $IW = none$; red square: $IW = MVIW$.

According to both Fig. 3 and Fig. 4, excluding small variations at the very last iterations, the same general trend expressed previously is maintained in all of the microarray datasets, leading us to conclude that MVIW is of little use for the aim of increasing feature subset stability.



(a) Colon (b) Leukemia (c) Prostate (d) Lung

(e) Breast (f) Melanoma (g) Parkinson

Fig. 5: Feature subset stability using RelievedF-RFE on the microarray data. Each plot shows the KI (vertical axis) over a set of RFE iterations (horizontal axis). Red squares: standard (unweighted) FS; blue asterisks: RLIW.

3.4 Suitability of the proposed RLIW method

This section presents the experiments performed using our novel RLIW method and the results obtained compared to the case where no IW is applied (Std-FS) in terms of feature subset stability and prediction accuracy. Our strategy introduces several improvements to the previous MVIW: first, a new approach for assessing the importance of each instance, as presented in Section 2; second, the weights are re-computed at each iteration of the RFE process, based on the *remaining* subset of features; third, the FS strategy of choice has been RelievedF-RFE, discarding SVM-RFE, given the high computational demands derived from the necessary optimization of the C parameter.

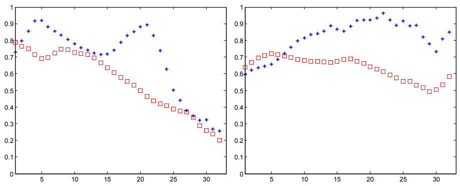
RLIW using microarray data. The experiments for microarray data have been developed using a *double 10-fold cross-validation* resampling strategy. Besides feature subset selection analysis, class predictions are obtained in the end by using a linear SVM, setting the C parameter to the value obtaining the best average balanced accuracy for the inner 10cv loop in a logarithmic scale. The final subset of features is the one reaching maximum stability, among those numbering less than 20% of the total number of features. The reported results obtained by RLIW measure the outer 10cv feature subset stability (KI measure) and are shown in Fig. 5; RLIW outperforms Std-FS across most RFE iterations for the Colon, Leukemia, Lung, Breast, Melanoma and Parkinson datasets. The Prostate dataset is an exception, for which we currently have no explanation beyond some particularity of the dataset.

	Colon	Leukemia	Prostate	Lung	Breast	Melanoma	Parkinson
Std-FS	0.82 ± 0.05 (22)	0.97 ± 0.02 (40)	0.94 ± 0.02 (5)	0.98 ± 0.01 (1026)	0.76 ± 0.05 (1026)	0.98 ± 0.02 (3)	0.78 ± 0.04 (1026)
RLIW-FS	0.79 ± 0.05 (22)	0.98 ± 0.02 (3)	0.92 ± 0.03 (1239)	0.97 ± 0.01 (19)	0.66 ± 0.05 (1026)	0.97 ± 0.02 (187)	0.68 ± 0.05 (923)

Table 1: Average balanced accuracies and their standard errors on the microarray datasets; feature subset size is shown in parentheses.

In addition, Table 1 presents the average BACs (and their standard errors) in predicting the class labels of the outer loop 10cv test set instances. A price is paid in terms of accuracy in exchange for the improvement in stability when using Breast and Parkinson data; in the rest of the problems, comparable accuracy is achieved.

RLIW using MRS data. The availability of a real test set allows to switch back to a standard 10 times 10-fold cross-validation (10x10cv) resampling strategy. This independent set has been used to properly validate the real difference in accuracy when using RLIW. The results, shown in Fig. 6, indicate that a similar performance can be achieved with more stable and smaller feature subsets. It is also apparent that the estimated 10x10cv and test prediction errors are in a better agreement using RLIW.



(a) STE (b) LTE

		10x10CV	Test
STE	Std-FS (36)	0.62 ± 0.01	0.67 ± 0.07
	RLIW-FS (17)	0.65 ± 0.01	0.68 ± 0.07
LTE	Std-FS (28)	0.62 ± 0.01	0.60 ± 0.08
	RLIW-FS (15)	0.65 ± 0.01	0.60 ± 0.08

(c) Balanced Accuracies

Fig. 6: a) and b) Feature subset stability over 10x10cv. Each plot shows the KI over the successive RFE iterations. Red squares: standard (unweighted) FS; blue asterisks: RLIW. The table in c) shows balanced accuracies and standard errors achieved by a linear SVM (number of selected features in parentheses).

4 Conclusions

The present work introduces RLIW, a new method for improving the stability of feature subset selection algorithms. Its suitability for medical practice has been assessed using data from two different environments: microarray gene expression and magnetic resonance spectroscopy of brain tumours. The reported results suggest a trade-off between prediction accuracy and feature subset stability. In many problems, a similar prediction performance figure is obtained but showing an increase in stability as measured by the Kuncheva index. In a few cases, however, this increase comes at the expense of a significant drop in prediction performance. We conjecture that, given the large dimensionality and the small sample sizes, it may well be that previous results, obtained without little concern for stability, are subject to large variability and thus rather optimistic in their evaluation of performance. Future research will investigate deeper this issue, and will evaluate the use of instance weighting into the learning algorithm itself.

Acknowledgments

This research was funded by Spanish MICINN R+D Plan TIN2009-13895-C02-01 project. The authors gratefully acknowledge former INTERPRET, eTUMOUR and HealthAgents European project partners. Data providers: Dr J. Capellades (Institut de Diagnòstic per la Imatge, Spain); Dr F.A. Howe and Professor J. Griffiths (St George's University of London, UK); Prof. A. Heerschap (RU); Prof. L. Stefanczyk and Dr J. Fortuniak (Uniwersytet Medycznyw Lodz, Poland); Dr J. Calvar (Fundación para la Lucha contra las Enfermedades Neurológicas de la Infancia, Argentina).

References

1. Kuncheva, L.I.: A stability index for feature selection. In: Procs. of the 25th IASTED International Multi-Conference. AIAP'07, ACTA Press (2007) 390–395
2. Somol, P., Novovičová, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(11) (2010) 1921–1939

3. Saeys, Y., Abeel, T., Peer, Y.: 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, Berlin, Heidelberg, Springer-Verlag (2008) 313–325
4. Loscalzo, S., Yu, L., Ding, C.: Consensus group stable feature selection. In: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2009) 567–576
5. Han, Y., Yu, L.: A Variance Reduction Framework for Stable Feature Selection. Statistical Analysis and Data Mining **5** (2012) 428–445
6. Julià-Sapé, M., Acosta, D., Mier, M., Arús, C., Watson, D.: A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA) **19** (2006) 22–33
7. González-Vélez, H., Mier, M., Julià-Sapé, M., et al.: Healthagents: Distributed multi-agent brain tumor diagnosis and prognosis. Journal of Applied Intelligence **30**(3) (2009) 191–202
8. Ranb, R.G.B., Navot, A., Tishby, N.: Margin based feature selection - theory and algorithms. In: Intl. Conf. on Machine Learning (ICML), ACM Press (2004) 43–50
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46**(1–3) (2002) 389–422
10. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(1) (1997) 273–324
11. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America **96**(12) (1999) 6745–6750
12. Golub, T.R., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537
13. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1**(2) (2002) 203–209
14. Lai, Y., Wu, B., Chen, L., Zhao, H.: A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics **20**(17) (2004) 3146–3155
15. Gordon, G.J., et al.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res **62** (2002) 4963–4967
16. van 't Veer, L.J., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**(6871) (2002) 530–536
17. Talantov, D., Mazumder, A., Yu, J.X., Briggs, T., Jiang, Y., Backus, J., Atkins, D., Wang, Y.: Novel genes associated with malignant melanoma but not benign melanocytic lesions. Clinical Cancer Research **11**(20) (2005) 7234–7242
18. Scherzer, C.R., et al.: Molecular markers of early parkinson's disease based on gene expression in blood. Proc. of the Natl. Acad. of Sciences **104**(3) (2007) 955–960
19. Opstad, K., Murphy, M., et al.: Differentiation of metastases from high-grade gliomas using short echo time ^{1}H spectroscopy. Journal of Magnetic Resonance Imaging **20**(2) (2004) 187–192
20. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: Proceedings of the IEEE 2010 International Conference on Pattern Recognition (ICPR). (2010) 3121–3124