Ensemble Logistic Regression for Feature Selection

Roman Zakharov and Pierre Dupont

Machine Learning Group,
ICTEAM Institute,
Université catholique de Louvain,
B-1348 Louvain-la-Neuve, Belgium
{roman.zakharov,pierre.dupont}@uclouvain.be

Abstract. This paper describes a novel feature selection algorithm embedded into logistic regression. It specifically addresses high dimensional data with few observations, which are commonly found in the biomedical domain such as microarray data. The overall objective is to optimize the predictive performance of a classifier while favoring also sparse and stable models.

Feature relevance is first estimated according to a simple t-test ranking. This initial feature relevance is treated as a feature sampling probability and a multivariate logistic regression is iteratively reestimated on subsets of randomly and non-uniformly sampled features. At each iteration, the feature sampling probability is adapted according to the predictive performance and the weights of the logistic regression. Globally, the proposed selection method can be seen as an ensemble of logistic regression models voting jointly for the final relevance of features.

Practical experiments reported on several microarray datasets show that the proposed method offers a comparable or better stability and significantly better predictive performances than logistic regression regularized with Elastic Net. It also outperforms a selection based on Random Forests, another popular embedded feature selection from an ensemble of classifiers.

Keywords: stability of gene selection, microarray data classification, logistic regression.

1 Introduction

Logistic regression is a standard statistical technique addressing binary classification problems [5]. However logistic regression models tend to over-fit the learning sample when the number p of features, or input variables, largely exceeds the number n of samples. This is referred to as the *small n large p* setting, commonly found in biomedical problems such as gene selection from microarray data.

A typical solution to prevent over-fitting considers an l_2 norm penalty on the regression weight values, as in ridge regression [10], or an l_1 norm penalty for

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 133–144, 2011.

the (Generalized) LASSO [20,16], possibly a combination of both, as in Elastic Net [22]. The l_1 penalty has the additional advantage of forcing the solution to be *sparse*, hence performing feature selection jointly with the classifier estimation.

Feature selection aims at improving the interpretability of the classifiers, tends to reduce the computational complexity when predicting the class of new observations and may sometimes improve the predictive performances [8,17]. The feature selection obtained with a LASSO type penalty is however typically unstable in the sense that it can be largely affected by slight modifications of the learning sample (*e.g.* by adding or removing a few observations). The stability of feature selection has received a recent attention [12,1] and the interested reader is referred to a comparative study of various selection methods over a number of high-dimensional datasets [11].

In this paper, we propose a novel approach to perform feature (*e.g.* gene) selection jointly with the estimation of a binary classifier. The overall objective is to optimize the predictive performance of the classifier while favoring at the same time sparse and stable models. The proposed technique is essentially an embedded approach [8] relying on logistic regression. This classifier is chosen because, if well regularized, it tends to offer good predictive performances and its probabilistic output helps assigning a confidence level to the predicted class.

The proposed approach nonetheless starts from a t-test ranking method as a first guess on feature relevance. Such a simple univariate selection ignores the dependence between features [17] but generally offers a stable selection. This initial feature relevance is treated as a feature sampling probability and a multivariate logistic regression model is iteratively reestimated on subsets of randomly, and non-uniformly, sampled features. The number of features sampled at each iteration is constrained to be equal to the number of samples. Such a constraint enforces the desired sparsity of the model without resorting on a l_1 penalty. At each iteration, the sampling probability of any feature used is adapted according to the predictive performance of the current logistic regression. Such a procedure follows the spirit of wrapper methods, where the classifier performance drives the search of selected features. However it is used here in a smoother fashion by increasing or decreasing the probability of sampling a feature in subsequent iterations. The amplitude of the update of the sampling probability of a feature also depends on its absolute weight in the logistic regression. Globally, this feature selection approach can be seen as an ensemble learning made of a committee of logistic regression models voting jointly for the final relevance of each feature.

Regularized logistic regression methods are briefly reviewed in section 2.1. Section 2.2 further details our proposed method for feature selection. Practical experiments of gene selection from various microarrays datasets, described in section 3, illustrate the benefits of the proposed approach. In particular, our method offers significantly better predictive performances than logistic regression models regularized with Elastic Net. It also outperforms a selection with Random Forests, another popular ensemble learning approach, both in terms of predictive performance and stability. We conclude and present our future work in section 4.

2 Feature Selection Methods

Ensemble learning has been initially proposed to combine learner decisions, which aggregation produces a single regression value or class label [7]. The idea of ensemble learning has also been extended to feature selection [8]. Approaches along those lines include the definition of a feature relevance from Random Forests [3] or the aggregation of various feature rankings obtained from a SVM-based classifier [1]. Those approaches rely on various resamplings of the learning sample. Hence, the diversity of the ensemble is obtained by considering various subsets of training instances. We opt here for an alternative way of producing diversity, namely by sampling the feature space according to a probability distribution, which is iteratively refined to better model the relevance of each feature.

In section 2.1, we briefly review logistic regression, which serves here as the base classifier. Section 2.2 further details the proposed approach of ensemble logistic regression with feature resampling.

2.1 Regularized Logistic Regression

Let $\mathbf{x} \in \mathbf{R}^p$ denote an observation made of p feature values and let $y \in \{-1, +1\}$ denote the corresponding binary output or class label. A logistic regression models the conditional probability distribution of the class label y, given a feature vector $\mathbf{x}$ as follows.

$$\mathrm{Prob}(y|\mathbf{x}) = \frac{1}{1 + \exp\left(-y(\mathbf{w}^T\mathbf{x} + v)\right)}, \tag{1}$$

where the weight vector $\mathbf{w} \in \mathbf{R}^p$ and intercept $v \in \mathbf{R}$ are the parameters of the logistic regression model. The equation $\mathbf{w}^T\mathbf{x} + v = 0$ defines an hyperplane in feature space, which is the decision boundary on which the conditional probability of each possible output value is equal to $\frac{1}{2}$.

We consider a supervised learning task where we have n i.i.d. training instances $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$. The likelihood function associated with the learning sample is $\prod_{i=1}^{n} \mathrm{Prob}(y_i|\mathbf{x}_i)$, and the negative of the log-likelihood function divided by n, sometimes called the *average logistic loss*, is given by

$$l_{avg}(\mathbf{w}, v) = \frac{1}{n} \sum_{i=1}^{n} f\left(y_i(\mathbf{w}^T\mathbf{x}_i + v)\right), \tag{2}$$

where $f(z) = \log\left(1 + \exp(-z)\right)$ is the *logistic loss function.*

A maximum likelihood estimation of the model parameters $\mathbf{w}$ and v would be obtained by minimizing (2) with respect to the variables $\mathbf{w} \in \mathbf{R}^p$ and $v \in \mathbf{R}$. This minimization is called the *logistic regression* (LR) problem. When the number n of observations is small compared to the number p of features, a logistic regression model tends to over-fit the learning sample. When over-fitting occurs many features have large absolute weight values, and small changes of those values have a significant impact on the predicted output.

The most common way to reduce over-fitting is to add a penalty term to the loss function in order to prevent large weights. Such a penalization, also known as regularization, gives rise to the l_2-regularized LR problem:

$$\min_{\mathbf{w},v} l_{avg}(\mathbf{w},v) + \lambda\|\mathbf{w}\|_2^2 = \min_{\mathbf{w},v} \frac{1}{n}\sum_{i=1}^{n} f\left(y_i(\mathbf{w}^T\mathbf{x}_i + v)\right) + \lambda\sum_{j=1}^{p} w_j^2. \tag{3}$$

Here $\lambda > 0$ is a regularization parameter which controls the trade-off between the loss function minimization and the size of the weight vector, measured by its l_2-norm.

As discussed in [15], the l_2-regularized LR worst case sample complexity grows at least linearly in the number of (possibly irrelevant) features. This result means that, to get good predictive performance, adding a feature to the model requires the inclusion of an additional learning example. For small n, large p problems the l_1-regularized LR is thus usually considered instead by replacing the l_2-norm $\|\cdot\|_2^2$ in (3) by the l_1-norm $\|\cdot\|_1$. This is a natural extension to the LASSO [20] for binary classification problems.

The benefit of the l_1-regularized LR is its logarithmic rather than linear sample complexity. It also produces *sparse* models, for which most weights are equal to 0, hence performing an implicit feature selection. However l_1-regularized LR is sometimes too sparse and tends to produce a highly unstable feature selection. A trade-off is to consider a mixed regularization relying on the Elastic Net penalty [22]:

$$\min_{\mathbf{w},v} l_{avg}(\mathbf{w},v) + \lambda\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)w_j^2 + \alpha|w_j|\right], \tag{4}$$

where $\alpha \in [0,1]$ is a meta-parameter controlling the influence of each norm. For high-dimensional datasets the key control parameter is usually still the l_1 penalty, with the l_2 norm offering an additional smoothing.

We argue in this paper that there is an alternative way of obtaining sparse and stable logistic regression models. Rather than relying on a regularization including an l_1 penalty, the sparsity is obtained by constraining the model to be built on a number of features of the same order as the number of available samples. This constraint is implemented by sampling feature subsets of a prescribed size. The key ingredient of such an approach, further detailed in section 2.2, is a non-uniform sampling probability of each feature where such a probability is proportional to the estimated feature relevance.

2.2 Ensemble Logistic Regression with Feature Resampling

The proposed feature selection is essentially an embedded method relying on regularized logistic regression models. Those models are built on small subsets of the full feature space by sampling at random this space. The sampling probability is directly proportional to the estimated feature relevance. The initial relevance of each feature is estimated according to a t-test ranking. Such a simple univariate ranking does not consider the dependence between features but is observed to be stable with respect to variations of the learning sample. This initial relevance index is iteratively refined as a function of the predictive performance of regularized logistic regression models built on resampled features. This procedure iterates until convergence of the classifier performance.

Our method relies on the l_2-regularized LR as estimated by the optimization problem (3). The sparsity is not enforced here with an l_1 penalty but rather by explicitly limiting the number of features on which such a model is estimated. The sample complexity result from [15] gives us a reasonable default number of features to be equal to the number of training examples n. Those n features could be drawn uniformly from the full set of p features (with $p \gg n$) but we will show the benefits of using a non-uniform sampling probability. We propose here to relate the sampling probability of a given feature to its estimated relevance.

Since our primary application of interest is the classification of microarray data, a t-test relevance index looks to be a reasonable choice as a first guess [21,8]. This method ranks features by their normalized difference between mean expression values across classes:

$$t_j = \frac{\mu_{j+} \; - \; \mu_{j-}}{\sqrt{\sigma_{j+}^2/m_+ \; + \; \sigma_{j-}^2/m_-}}, \tag{5}$$

where μ_{j+} (respectively μ_{j-}) is the mean expression value of the feature j for the m_+ positively (respectively m_- negatively) labeled examples, and σ_{j+}, σ_{j-} are the associated standard deviations. The score vector $\mathbf{t}$ over the p features is normalized to produce a valid probability distribution vector **prob**. We note that there is no need here to correct for multiple testing since the t-test is not used to directly select features but to define an initial feature sampling probability.

At each iteration the learning sample is split into training (80%) and validation (20%) sets. Next, a subset of n features is drawn according to **prob** and a $l2$-regularized LR model is estimated on the training data restricted to those features. The resulting classifier is evaluated on the validation set according to its balanced classification rate BCR, which is the average between specificity and sensitivity (see section 3.2).

The BCR performance of the current model is compared to the average $\overline{BCR}$ (initially set to 0.5) obtained for all models built at previous iterations. The current model quality is estimated by $\log(1 + BCR - \overline{BCR})$. The relative quality of the current model is thus considered positive (resp. negative) if its performance is above (resp. below) average.

Finally, the probability vector **prob** controlling the sampling of features at the next iteration is updated according to the relative model quality and its respective weight vector $\mathbf{w}$. The objective is to favor the further sampling of important features (large weight values) whenever the current model looks good and disfavor the sampling of non-important features (small weight values) when the current model looks poor. This process is iterated until convergence of the classification performance. The net result of this algorithm summarized below is the vector **prob** which is interpreted as the final feature relevance vector.

Algorithm 1. Ensemble Logistic Regression with Feature Resampling

```
Algorithm ELR
Input: A learning sample $\mathbf{X} \in \mathbf{R}^{n \times p}$ and class labels $\mathbf{y} \in \{-1, 1\}^n$
Input: A regularization parameter $\lambda$ for estimating a $l2$-LR model (3)
Output: A vector $\mathbf{prob} \in [0, 1]^p$ of feature relevance

Initialize prob according to a $t$-test ranking
$\overline{BCR} \leftarrow 0.5$        // Default initialization of the average BCR

repeat
    Randomly split X into TRAINING (80%) and VALIDATION (20%)
    Draw $n$ out of $p$ features at random according to prob
    $(\mathbf{w}, v) \leftarrow$ a $l2$-LR model $M$ learned on TRAINING restricted to $n$ features
    Compute BCR of $M$ on VALIDATION
    $quality \leftarrow \log(1 + BCR - \overline{BCR})$
    // Update the feature relevance vector
    foreach $j$ among the $n$ sampled features do
        $\mathbf{prob}_j \leftarrow \frac{1}{Z}\left(\mathbf{prob}_j + quality \cdot {\mathbf{w}_j}^{2 \cdot \mathrm{sign}(quality)}\right)$
        // $Z$ is the normalization constant to define a distribution
    Update average $\overline{BCR}$
until no significant change of $\overline{BCR}$ between consecutive iterations;
return prob
```

3 Experiments

3.1 Microarray Datasets

We report here practical experiments of gene selection from 4 microarray datasets. Table 1 summarizes the main characteristics of those datasets: the number of samples, the number of features (genes) and the class ratios.

The classification task in DLBCL (diffuse large B-cells) is the prediction of the tissue type [18]. Chandran [4] and Singh [19] are two datasets related to prostate cancer and the task is to discriminate between tumor or normal samples. The Transbig dataset is part of a large breast cancer study [6]. The original data measures the time to metastasis after treatment. We approximate this task here (for the sake of considering an additional dataset) by considering a time threshold of 5 years after treatment. Such a threshold is commonly used as a critical value in breast cancer studies. The question of interest is to discriminate between patients with or without metastasis at this term, and hence reduces to a binary classification problem. We focus in particular on ER-positive/HER2-negative patients, which form the most significant sub-population as reported in [13].

Table 1. Characteristics of the microarray datasets

Dataset	Samples (n)	Features (p)	Class Priors
DLBCL	77	7129	25%/75%
Singh	102	12625	49%/51%
Chandran	104	12625	17%/83%
Transbig	116	22283	87%/13%

3.2 Evaluation Metrics

The main objective is to assess the predictive performance of a classifier built on the selected genes. The performance metric is estimated according to the *Balanced Classification Rate*:

$$BCR = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right), \tag{6}$$

where $TP(TN)$ is the number of correctly predicted positive (negative) test examples among the P positive (N negative) test examples. BCR is preferred to the classification accuracy because microarray datasets often have unequal class prior, as illustrated in Table 1. BCR is the average between specificity and sensitivity and can be generalized to multi-class problems more easily than ROC analysis.

To further assess the quality of feature (= gene) selection methods, we evaluate the stability of the selection on k resamplings of the data. The Kuncheva index [12] measures to which extent k sets of s selected features share common features.

$$K\left(\{\mathbf{S}_1, \ldots, \mathbf{S}_k\}\right) = \frac{2}{k\left(k-1\right)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{\mid \mathbf{S}_i \cap \mathbf{S}_j \mid - \frac{s^2}{p}}{s - \frac{s^2}{p}}, \tag{7}$$

where p is the total number of features, and $\mathbf{S}_i$, $\mathbf{S}_j$ are two gene lists built from different resamplings of the data. The s^2/p term corrects a bias due to chance of selecting common features among two sets chosen at random. The Kuncheva index ranges within (-1,1] and the greater its value the larger the number of common features across the k gene lists.

3.3 Experimental Methodology

We report experimental performances of the gene selection approach introduced in section 2.2 and referred to as ELR. A simple variant, denoted by ELR_WOTT, uses a uniform distribution over the p genes to initialize the sampling probability distribution **prob**, instead of the t-test values. We also consider, as an additional baseline denoted TTEST, a direct ranking of the genes according to the t-test statistics without any further refinement (hence reducing ELR to its initialization).

A further competing approach is ENET: a gene selection based on the absolute values of the feature weights estimated from a logistic regression regularized with Elastic Net. In such an approach, the sparsity is controlled by the regularization constants λ and α (see equation (4)). We choose $\alpha = 0.2$ as in the original work on microarray classification from the authors [22], and let λ vary in the range $[10^{-6}, 10]$ to get more or fewer selected features. In contrast, in the ELR method, which uses a $l2$-regularized logistic loss, λ is fixed to a default value equal to 1. In this case, the sparsity results from the limited number of sampled features and the final result is a ranking of the full set of features according to the **prob** vector.

In contrast to the ELR method, which relies on various resamplings from the set of *features*, alternative methods use several bootstrap samples from the set of *training examples*. We report comparative results with BoRFE a bootstrap extension to RFE [1]. This method is similar to BoLASSO [2] but tailored to classification problems. Following [1], we rely on 40 bootstrap samples while discarding 20 % of features at each iteration of RFE.

Random Forests (RF) are another competing approach to define a relevance measure on genes [3]. Here, a key control parameter is the number of trees considered in the forest. Preliminary experiments (not reported here) have shown that the predictive performance obtained with RF is well stabilized with 1,000 trees. The stability of the gene selection itself can be improved by considering a larger number of trees ($\approx 5,000$) however resulting sometimes in a lower BCR. Hence we stick to $\approx 1,200$ trees on DLBCL, $\approx 2,000$ trees on Singh and Chandran, and $\approx$ 2,500 trees on Transbig. Those numbers also happen to be of a similar order of magnitude as the number of iterations used by the ELR method to converge. We thus compare two different selection methods from approximately the same number of individual models. The second RF meta-parameter is the number of features selected at random when growing each node of the trees in the forest. To be consistent with the ELR method, we choose to sample n features at each node. We note that, given the characteristics of the datasets under study (see table 1), the n value tends to be close to $\sqrt{p}$, which is a common choice while estimating a RF [8].

All methods mentioned above produce ranked gene lists from expression values as measured by microarrays. Specific gene lists are obtained by thresholding such lists at a prescribed size. We report performances for various list sizes from 512 genes down to 4 genes. We consider 200 independent sub-samplings without replacement forming binary splits of each dataset into 90% training and 10% tests. The stability of the gene selection is evaluated according to the Kuncheva index over the 200 training sets. The predictive performance of classifiers built on the training sets from those genes is evaluated and averaged over the 200 test sets. To compare predictive results only influenced by the gene selection, we report the average BCR of $l2$-regularized LR classifiers, no matter which selection method is used.

3.4 Results

Figures 1 and 2 report the BCR and stability performances of the various methods tested on the 4 datasets described in table 1. The gene selection stability of the ELR method is clearly improved when using a t-test to initialize the feature sampling probability rather than a uniform distribution (ELR_WOTT). This result illustrates that a non-uniform feature sampling related to the estimated feature relevance is beneficial for the selection stability with no significant influence on the predictive performance. ELR is also more stable than RF on 3 out of 4 datasets while it offers results comparable to ENET and BoRFE in this regard.

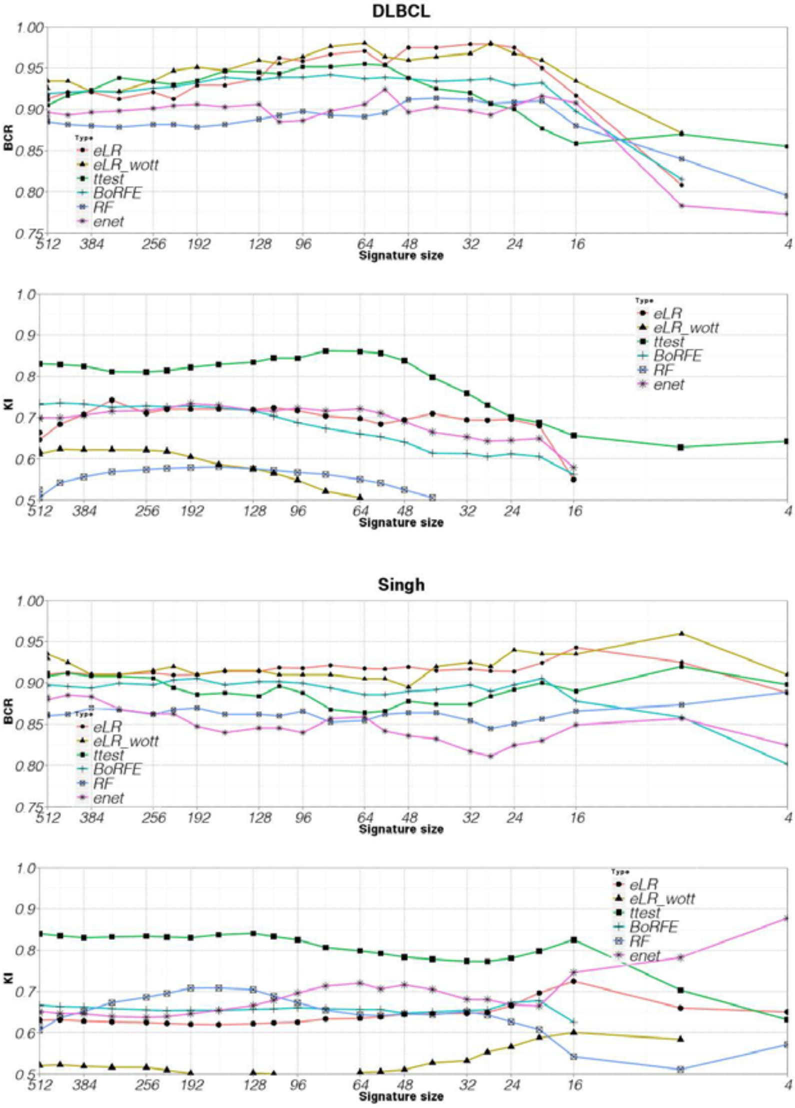

Fig. 1. Classification performance (BCR) and signature stability (Kuncheva index) of the competing methods on the **DLBCL** and **Singh** datasets

The ELR method outperforms, generally significantly, its competitors in terms of predictive performance. To support this claim, we assess the statistical significance of the differences of average BCR obtained with the ELR method and each of its competitors. We resort on the corrected resampled t-test proposed in [14] to take into account the fact that the various test folds do overlap. The significance is evaluated on the smallest signature size which show the highest BCR value: 24 genes for DLBCL, 16 genes for Singh, 28 genes for Chandran and 160 genes for Transbig. Table 2 reports the p-values of the pairwise comparisons between ELR and its competitors. Significant results (p-value < 0.05) are reported in bold in this table. ELR clearly outperforms ENET on all datasets and RF on 3 out of 4 datasets. It offers better performances than BoRFE and TTEST on all datasets, and significantly on DLBCL and Transbig respectively.

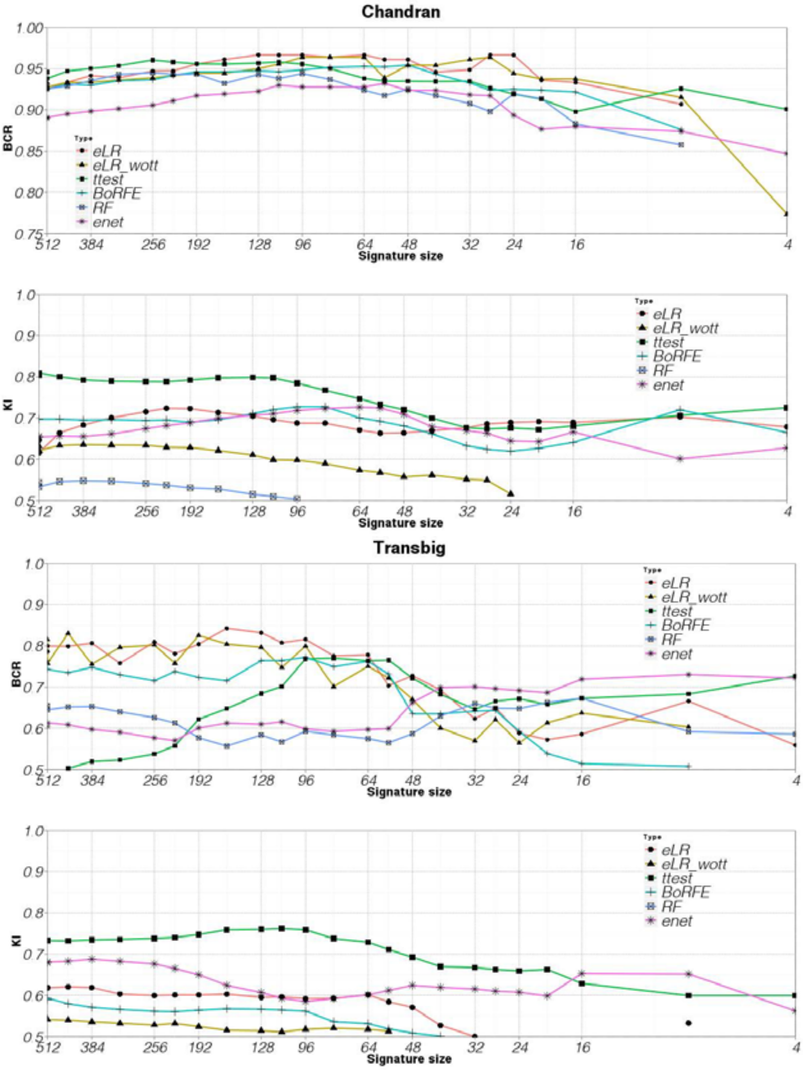

Fig. 2. Classification performance (BCR) and signature stability (Kuncheva index) of the competing methods on the **Chandran** and **Transbig** datasets

Table 2. Pairwise comparison of the average BCR obtained of ELR and its competitors. Reported results are p-values computed according to the corrected resampled t-test proposed in [14].

ELR vs.	DLBCL	Singh	Chandran	Transbig
Elastic Net	**0.039**	**0.001**	**0.043**	**0.033**
Random Forests	**0.023**	**0.007**	0.086	**0.005**
Boost. RFE	**0.042**	0.116	0.091	0.056
t-test	0.089	0.135	0.092	**0.046**
# genes	24	16	28	160

4 Conclusion and Future Work

We propose a novel feature selection method tailored to high dimensional datasets. The selection is embedded into logistic regression (LR) with non-uniform feature sampling. The sampling distribution of features is directly proportional to the estimated feature relevance. Such relevance is initialized with a standard t-test and further refined according to the predictive performance and weight values of LR models built on the sampled features. Experiments conducted on 4 microarray datasets related to the classification of tumor samples illustrate the benefits of the proposed approach in terms of predictive performance and stability of the gene selection.

Our future work includes a more formal analysis of the sampling probability update rule. On a practical viewpoint, the initial feature relevance could also be adapted according to some prior knowledge on some genes *a priori* believed to be more relevant. Such an approach would be an interesting alternative to the partially supervised gene selection method proposed in [9].

References

1. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26, 392–398 (2010)
2. Bach, F.R.: Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th International Conference on Machine Learning, pp. 33–40. ACM (2008)
3. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
4. Chandran, U.R., Ma, C., Dhir, R., Bisceglia, M., Lyons-Weiler, M., Liang, W., Michalopoulos, G., Becich, M., Monzon, F.A.: Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC Cancer 7(1), 64 (2007)
5. Cox, D.R., Snell, E.J.: Analysis of binary data. Monographs on statistics and applied probability. Chapman and Hall (1989)

6. Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., D'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Klijn, J., Foekens, J., Cardoso, F., Piccart, M., Buyse, M., Sotiriou, C.: Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. Clinical Cancer Research 13(11), 3207–3214 (2007)
7. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
8. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (eds.): Feature Extraction. Foundations and Applications. Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer (2006)
9. Helleputte, T., Dupont, P.: Feature Selection by Transfer Learning with Linear Regularized Models. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS, vol. 5781, pp. 533–547. Springer, Heidelberg (2009)
10. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67 (1970)
11. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems 12, 95–116 (2007), doi:10.1007/s10115-006-0040-8
12. Kuncheva, L.I.: A stability index for feature selection. In: Proceedings of the 25th International Multi-Conference Artificial Intelligence and Applications, pp. 390–395. ACTA Press, Anaheim (2007)
13. Li, Q., Eklund, A.C., Juul, N., Haibe-Kains, B., Workman, C.T., Richardson, A.L., Szallasi, Z., Swanton, C.: Minimising immunohistochemical false negative er classification using a complementary 23 gene expression signature of er status. PLoS ONE 5(12), e15031 (2010)
14. Nadeau, C., Bengio, Y.: Inference for the generalization error. Machine Learning 52, 239–281 (2003)
15. Ng, A.Y.: Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In: Proceedings of the Twenty-First International Conference on Machine Learning (ICML), vol. 1, pp. 78–85 (2004)
16. Roth, V.: The generalized LASSO. IEEE Transactions on Neural Networks 15(1), 16–28 (2004)
17. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
18. Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, A., Mesirov, J.: Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8, 68–74 (2002)
19. 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., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209 (2002)
20. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)
21. Witten, D.M., Tibshirani, R.: A comparison of fold-change and the t-statistic for microarray data analysis. Stanford University. Technical report (2007)
22. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301–320 (2005)