Classification of high dimensional data using LASSO ensembles

Daniel Urda
Universidad de Málaga, Andalucía Tech
ETSI Informática (España)
Instituto de Investigación Biomédica de Málaga (IBIMA)
Marbella International University Center, MIUC (España)
Email: durda@lcc.uma.es

Leonardo Franco
and José M. Jerez
Universidad de Málaga, Andalucía Tech
Dpto. de Lenguajes y Ciencias de la Computación (España) e
Instituto de Investigación Biomédica de Málaga (IBIMA)
Email:{lfranco,jja} @lcc.uma.es

***Abstract*—The estimation of multivariable predictors with good performance in high dimensional settings is a crucial task in biomedical contexts. Usually, solutions based on the application of a single machine learning model are provided while the use of ensemble methods is often overlooked within this area despite the well-known benefits that these methods provide in terms of predictive performance. In this paper, four ensemble approaches are described using LASSO base learners to predict the vital status of a patient from RNA-Seq gene expression data. The results of the analysis carried out in a public breast invasive cancer (BRCA) dataset shows that the ensemble approaches outperform statistically significant the standard LASSO model considered as baseline case. We also perform an analysis of the computational costs involved for each of the approaches, providing different usage recommendations according to the available computational power.**

I. Introduction

Machine learning (ML) models are nowadays very frequently applied in biomedical-related areas [1]–[3], as they normally outperform univariable predictors [4]. However, developing multivariable predictors using high dimensional data becomes an issue since ML models face the large-*p*-small-*n* problem (thousands of variables and a few hundreds of samples usually available), thus resulting in a highly negative impact on the predictor's performance.

In this sense, the amount of studies published describing the use of one type of ML model in predictive modeling has grown considerable in recent years. To cite just a few of them, for example, a l_1-regularization model enriched with biological knowledge was recently proposed in [5]. Before, Cui & Wang [6] proposed an approach that combines a linear model with l_1-regularization constraint (LASSO) and a neural network (NN) initialized with random weights. Different support vector machines (SVM) variants were also proposed in [7] to analyze high dimensional data. Moreover, a fast k-nearest neighbour (kNN) implementation or a sparse linear discriminant analysis (LDA) by thresholding were described in [8] and [9] respectively. However, none of these works considered the multi-view approach provided by ensemble methods where the optimal solution is approximate by a consortium of multiple individual ML models, usually boosting the overall performance [10], [11].

One application of an ensemble method in high dimensional data was provided by Do et al. in [12] using random forest (RF) to analyze high dimensional data as an ensemble of decision trees base learners. Nevertheless, decision trees are not suitable enough to deal with large number of input variables, usually requiring to combine them with a feature dimensionality reduction technique. On the other hand, Song et al. [13] proposed an ensemble of generalized linear models (GLM) with a forward selection procedure to choose the top-P most significant genes to be used within each base learner (P being a parameter of the model), thus imposing a constrain in the dimensionality of the input space to avoid overfitting issues of linear models. More recently, Wang et al. [14] proposed an ensemble of LASSO models to predict credit scoring within a dataset containing 80 variables after including some original variables' transformation. Although this work presented an ensemble of LASSO base learners suitable for high dimensional dataset analysis, in reality this work used the proposed model in a relatively low-dimension data (they use only 11 original variables that were incremented up to 80 through derivation and transformation techniques).

Therefore, our work aims to test the predictive performance of an ensemble of LASSO models using real high dimensional data consisting of RNA-Seq gene expression profiles obtained from cancer patients. In concrete, our ensemble approach follows data diversification procedures suggested in [15], [16] in order to create different views of the data and individually estimate a LASSO base learner to each single view. Further, several aggregation procedures are proposed to combine the individual predictions. In particular, we applied a simple aggregation procedure such as weighted average, which has been proven to outperform a majority voting approach [17]. In addition, three other more sophisticated approaches based on ensembles of linear models with some kind of regularization are also used.

The rest of the paper is organized as follows. Section II contains a detailed description of the data set used in the analysis as well as the baseline and ensemble approaches used. Section III presents the validation strategy followed to estimate the predictive performance of the models. Finally, Sections IV and V show the results obtained with the proposed ensemble

methods as well as some conclusions and posibble extensions to this research work.

II. Materials and Methods

The dataset used in this analysis consists of N=1212 samples of records containing P=20021 variables for describing each sample. The data correspond to patients linked to breast invasive adenocarcinoma (BRCA) for whom tissue sample was sequenced to finally obtain the RNA-Seq gene expression profile. In other words, for a given patient this data set will contain a row of 20021 variables where each of them correspond to the expression level of a certain gene. The complete data set, after applying pre-processing procedures for batch correction and RSEM normalization [18], is freely available and can be downloaded at The Cancer Genome Atlas (TCGA) website[1]. Additionally, we first removed those genes that do not show any expression across the sample, as they do not add predictive value, and we performed a logarithmic (log_2) transformation of the expression levels to approximate them to a normal distribution. In terms of predictive modeling, this analysis aims to predict the vital status of a given patient (0 = "$alive$", 1 = "$dead$") at a fixed time t from the gene expression profile of a patient, thus being a binary classification problem. The data is highly imbalanced since the class proportion consists of 1013 controls (alive) and only 199 cases (dead).

Regarding the methods considered within the analysis, our proposal follows an ensemble approach to combine predictions of several LASSO base learner models, using a standard LASSO model as baseline for comparison. Next, a more detailed description of the proposed methods is provided.

A. Standard LASSO

LASSO is a widely known model [19] that essentially consists of a simple linear model combined constraint with an l_1-penalty term to the objective function. Let us assume our data set is represented as $D = \{\boldsymbol{x_i}, y_i\}$, with $i \in \{1..N\}$ samples, $\boldsymbol{x_i}$ representing the vector of P genes describing the i-th sample, and y_i being the class (target) label. Then, Eq. 1 shows the objective function that is minimized under the LASSO approach for the case of a binary classification problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_i - F_{sig}(\boldsymbol{\beta x_i}))^2 + \lambda \sum_{j=1}^{P} |\beta_j| \qquad (1)$$

where the function F_{sig} represents the sigmoid function and is defined as follows:

$$F_{sig}(x) = \begin{cases} 1, & \text{if } \frac{1}{1+e^{-x}} \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

The LASSO model tries to set as many coefficients (β_j) as possible to zero unless a certain gene x_j is really important to drive the predictions right. LASSO models have been previously shown to work well in the large-p-small-n scenario being able to overcome overfitting issues. The amount of regularization applied is controlled by the hyper-parameter λ which takes values in the $(0, 1)$ range. When λ takes larger values, the L_1-penalty term in Eq. 1 has a higher incidence in the whole objective function and, therefore, less genes will be retained by the model. The value of the λ hyper-parameter is learned from data through a cross-validation process.

[1]https://cancergenome.nih.gov/

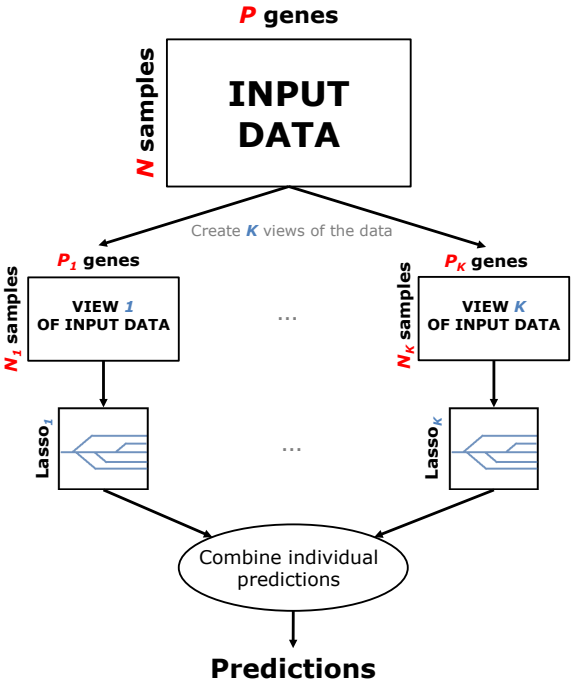

Fig. 1. Ensemble approach for a given input dataset of N samples and P genes.

B. Ensemble approach

This paper proposes to use an ensemble approach in a similar way to the one applied in Random Forests [16] to analyze RNA-Seq gene expression data. In contrast to this well-known machine learning model, this work uses a standard LASSO as base learner model, a model that will be referred as *RandomLasso* from this moment onwards. Ensembles methods such as Random Forest try to take advantage of several base learners trained on similar instances of the data, thus having different views of the problem, to finally combine the individual predictions of the learners to provide an overall prediction. In high dimensional data sets as the one being used within this work, view instances of the original data would be even more different and authors expect to see that the ensemble advantages are increased outperforming simpler models like the standard LASSO.

Figure 1 shows the architecture of the overall ensemble approach used in this work. N samples and P genes from the input dataset are considered, and two main aspects can be highlighted:

1) Base learners: The ensemble is composed by K base learners, LASSO in this case, each of them being individually fitted to an specific view of the original input data. These views are created as explained in [10] through manipulation of the training samples as well as the input features. In concrete, and due to the few number of samples N typically available in the large-p-small-n scenario, the complete dataset is sub-sampled K times in order to produce K views with at least 85% of the total number of samples N. Furthermore, a similar sub-sampling procedure is applied to have different subset of genes within the K views in such a way that the number of genes present on a single view will be around $2 * \sqrt{P}$. This last sub-sampling procedure is more aggressive and follows a similar strategy as the one applied in Random Forest [16], thus ending with K views of the input data that are very different one from the other. In addition, each single view created is a highly reduced version of the initial problem, what basically means that each individual problem is more manageable and could be easily computed by simple machine learning models such as LASSO.

2) Aggregation of base learners: In order to obtain an overall prediction using the proposed ensemble approach, it is required to somehow combine the individual predictions provided by each base learner. Simple procedures based on majority votes or averaging predictions are widely used by the machine learning community. However, in this paper we propose the use of some more sophisticated procedures to build a meta-model as follows:

- *Weighted Average*. This meta-model could be seen as an extended version of a simple aggregation procedure such as the average. This procedure uses the individual accuracy measured by the AUC (area under the ROC curve) obtained from each base learner in such a way that those base learners with higher AUC will have higher weight in the average calculation.
- *LASSO*. This meta-model is a linear model with l_1-regularization. This model will be more suitable to combine individual predictions of K base learners when K is a high number. The LASSO meta-model will only retain base learners that are useful in terms of predictive accuracy at the same time that it will get rid of those base learners that have no incidence in predicting the event of interest.
- *Ridge*. This meta-model is similar to the previous one but in this case the linear model includes a l_2-penalty term instead of the l_1. This means that this meta-model will not get rid of useless base learners, although it will assign them a tiny coefficient so that their predictions will have almost no incidence in predicting the event of interest.
- *Elastic Net*. The last meta-model is a linear model with a combined l_1 and l_2-regularization. It is something intermediate between the two previous ones, thus very poor base learners will be removed by this meta-model and, therefore, not used to compute final predictions. On the other hand, not very good base learners will be assigned with a tiny coefficient so that their predictions have little incidence in predicting the event of interest.

III. Validation Strategy

The validation strategy consists of 100 repetitions of 10-fold balanced cross-validation schemes to test the goodness of our ensemble approaches with respect to the baseline model. In this sense, each repetition will divide the original input data into 10 non-overlapping folds of equal sizes keeping the original class proportion within each fold. Moreover, every repetition will contain a fold partitioning completely different to other repetitions to avoid biasing our analysis due to the use of an specific partitioning of the input data.

Usually, this validation strategy iterates over the number of folds created (10 in this case), where on each iteration 9 folds are chosen to train or fit the model (thus 90% of the data) and the remaining outer fold (10% of the data) is used to test the performance of the trained model. However, this simple scheme cannot be used straightforward if we expect to make a fair comparison between the baseline model and our ensemble approaches. For a better understanding of the problem here, let us suppose that we use the same 9 folds used by our baseline LASSO to train K base learners from our ensemble approach. These K base learners trained could then be used to individually get predictions for each sample within the 9 folds. At this point, predictions must be combined to produce the final predictions by estimating some of the meta-models proposed in our ensemble approach. The question here would be: which part of the data could be used to estimate the meta-model? The remaining outer fold cannot be used since it is purely linked to unseen future data, meaning that it can only be used to measure the performance of the final model, in our case our ensemble including both base learners and meta-model. On the other hand, using the same 9 folds to estimate the meta-model would provide us over-optimistic results that will not be replicated in the outer fold left purely for testing. The reason is that base learners and the meta-model would be trained on the same samples and, therefore, the meta-model will favor the base learners with the lowest training error. Therefore, the 9 folds used to originally estimate the model should be somehow divided so that base learners are estimated in a subset of the 9 folds and the meta-model is estimated in the remaining subset.

Figure 2 shows a graphical representation of the 10-fold partitioning of one out of the 100 repetitions performed in the analysis. According to the explanation provided before, on a single iteration one of the folds (10% of the input data represented in green) is left apart and is only used to test the performance of the baseline model or the ensemble approach. Then, within the samples contained in the remaining 9 folds, a holdout strategy is applied in such a way that 72% of the original input data is used to estimate the K base learners (represented in blue) and 18% of the data is used to estimate the meta-model (represented in red). The blue and red blocks appear together in the image just for visualization purpose and for an easier understanding, although in reality the holdout strategy performs a random split. Logically, this

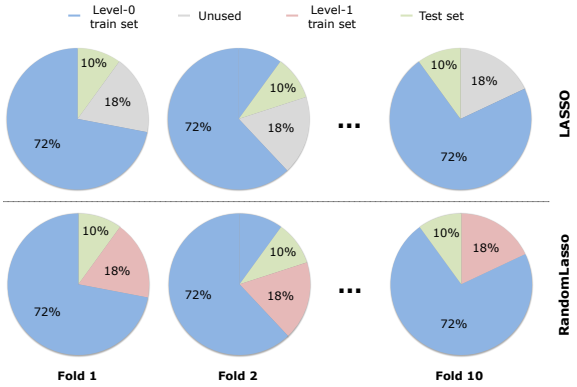

Fig. 2. Validation strategy used within the analysis. The baseline model and our ensemble approaches use the same level-0 train set (blue) to estimate the models and the same test set (green) to test the performance of the fitted models. Level-1 train set (red) is only used by our ensemble approaches to estimate the meta-model that combines predictions of individual base learners.

procedure could be improved by performing internal cross-validation within the 90% of the data instead of doing a simple holdout in order to build a more robust meta-model. However, authors decided to go with the holdout option described to keep the experimental design simpler. Furthermore, exactly the same partitioning described before was shared and used to estimate the baseline model (standard LASSO). In this sense, the 18% of the data used by the ensemble approach to estimate the meta-model remains here unused (represented in gray). In other words, LASSO is fitted to 72% of the data and tested in the outer 10%. Therefore, a fair comparison of the performance measure between LASSO and our ensemble approaches will be carried out.

IV. Results

All experiments carried out within this analysis used the strategy described in Section III, i.e., 100 repetitions of 10-fold cross-validation. The estimation of both baseline model and ensemble approaches were done under the R software and using the package "glmnet" [20] which already performs a nested cross-validation to learn the regularization parameter λ. As mentioned before, since the BRCA dataset is highly imbalanced, the Area Under the Curve (AUC) is used to measure the goodness of fit of a given model.

With respect to the ensemble approaches, a hyper-parameter K has to be chosen in order to specify the number of individual LASSO base learners that will be fitted within the ensemble. Ideally, this hyper-parameter should be learned from the data in such a way that the selected value minimizes the error in out-of-bag samples (e.g. through nested cross-validation). However, finding the optimal value of the hyper-parameter K is beyond the scope of this paper which aims to test the benefits of ensemble approaches compared to standard machine learning models for the analysis of high dimensional data. Therefore, some empirical values were chosen ($K \in \{5, 10, 50, 100, 500, 1000\}$).

Table I shows the raw results obtained in this analysis. In concrete, it shows the results for the baseline model (LASSO) and the 4 ensemble approaches proposed ($Random - WeightedAVG$, $RandomLasso - Lasso$, $RandomLasso - Ridge$, $RandomLasso - Elnet$), each of them considering different number of base learners. In addition, the table provides the average AUC obtained using each proposed model with the 95% confidence intervals (CI), as well as some extra information about the average number of genes retained by LASSO and the average time (in minutes) required to estimate the model. As shown in [21], the 95% CI provided should be taken carefully as they proved that there is no unbiased estimator of the variance of k-fold cross-validation, thus possibly representing over-optimistic results. For the ensemble approaches, the average number of genes showed represents the average number of retained genes across the K LASSO base learners. Regarding the time needed to compute the models, it represents an estimation since this analysis was run in a high performance cluster, meaning that this measure will be highly influenced by the characteristic of each individual node of the cluster as well as the load balance of the node where each experiment was being physically executed. Thus, the estimation of computational times was done taken an average value for a single computation of an ensemble model and then multiplying this value by K. Furthermore, the table shows the resulting p-value obtained by applying a Wilcoxon paired signed rank test [22]–[24] to test the statistical significance of the obtained AUC of the ensemble approaches with respect to the baseline LASSO model.

In terms of AUC, it can be clearly seen the existence of at least one configuration per ensemble approach that outperforms a standard LASSO model. In the worst case scenario, both $RandomLasso - Lasso$ and $RandomLasso - Elnet$ achieve an AUC of 0.65, that is 0.01 points slightly above the standard LASSO. On the other hand, both $RandomLasso - WeightedAVG$ and $RandomLasso - Ridge$ provide the best case scenarios pushing the AUC up to 0.67 and 0.68 respectively. Moreover, these improvements turned out to be statistically significant according to Wilcoxon signed rank test in most of the configurations analyzed. Regarding the impact of the number of base learners used within the ensemble approaches (K) in the predictive performance, Figure 3 shows that incrementing the number of base learners has a positive tendency in all the ensemble approaches, although the impact diminishes as soon as K grows. Furthermore, in the analyzed dataset setting $K = 100$ could be considered a save threshold to guarantee a better performance of any of the ensemble approaches, i.e. AUCs of 0.641, 0.644, 0.667 and 0.671 from the worst (red line) to the best (blue line) ensemble approach, in contrast to the AUC of 0.637 obtained by the standard LASSO.

With respect to the number of genes retained by each model, it can be highlighted that the ensemble approaches use less genes than the baseline model (around 90 genes in most of them compared to 218 genes retained by the baseline). However, this must be taken carefully since this measure

TABLE I

AVERAGE TEST DATA RESULTS FOR THE BASELINE (STANDARD LASSO) AND OUR PROPOSED ENSEMBLE MODELS AFTER 100 REPETITIONS OF 10-FOLD CROSS-VALIDATION. THE NUMBER OF BASE LEARNERS USED IN THE ENSEMBLE APPROACHES, THE AREA UNDER THE CURVE, THE AVERAGE NUMBER OF SELECTED GENES, THE AVERAGE TIME IN MINUTES NEEDED FOR THE ANALYSIS, AND THE LEVEL OF SIGNIFICANCE WITH RESPECT TO THE BASELINE (* P-VALUE<=0.05, ** P-VALUE<=0.01, *** P-VALUE<=0.001) ARE SHOWN FOR EACH CONSIDERED MODEL.

Model	K	AUC	#genes	time (mins.)	Wilcoxon signed rank test
LASSO	-	0.64 [0.60, 0.67]	218.52 ± 39.97	12.81 ± 2.96	-
RandomLasso - Weighted AVG	5	0.64 [0.6, 0.67]	90.68 ± 10.86	52.12 ± 8.80	0.75
RandomLasso - Weighted AVG	10	0.65 [0.61, 0.68]	91.20 ± 7.57	104.24 ± 17.61	$5.79*10^{-5}$ (***)
RandomLasso - Weighted AVG	50	0.67 [0.65, 0.69]	91.06 ± 3.50	521.20 ± 88.03	$<2.2*10^{-16}$ (***)
RandomLasso - Weighted AVG	100	0.67 [0.65, 0.69]	91.47 ± 3.08	1042.39 ± 176.06	$<2.2*10^{-16}$ (***)
RandomLasso - Weighted AVG	500	0.67 [0.65, 0.69]	91.60 ± 2.57	5211.95 ± 880.28	$<2.2*10^{-16}$ (***)
RandomLasso - Weighted AVG	1000	0.67 [0.65, 0.69]	91.88 ± 2.59	10423.90 ± 1760.56	$<2.2*10^{-16}$ (***)
RandomLasso - Lasso	5	0.63 [0.60, 0.66]	90.68 ± 10.86	52.13 ± 8.80	0.13
RandomLasso - Lasso	10	0.63 [0.60, 0.66]	91.20 ± 7.57	104.27 ± 17.61	0.12
RandomLasso - Lasso	50	0.64 [0.60, 0.67]	91.06 ± 3.49	521.35 ± 88.03	0.55
RandomLasso - Lasso	100	0.64 [0.60, 0.67]	91.47 ± 3.08	1042.69 ± 176.06	0.07
RandomLasso - Lasso	500	0.64 [0.61, 0.68]	91.60 ± 2.57	5213.47 ± 880.29	$4.84*10^{-3}$ (**)
RandomLasso - Lasso	1000	0.65 [0.61, 0.68]	91.88 ± 2.59	10426.94 ± 1760.59	$8.2*10^{-5}$ (***)
RandomLasso - Ridge	5	0.64 [0.61, 0.68]	90.68 ± 10.86	52.14 ± 8.80	$9.48*10^{-3}$ (**)
RandomLasso - Ridge	10	0.65 [0.62, 0.67]	91.20 ± 7.57	104.27 ± 17.61	$8.87*10^{-7}$ (***)
RandomLasso - Ridge	50	0.66 [0.63, 0.68]	91.06 ± 3.49	521.36 ± 88.03	$5.26*10^{-15}$ (***)
RandomLasso - Ridge	100	0.66 [0.64, 0.68]	91.47 ± 3.08	1042.72 ± 176.06	$<2.2*10^{-16}$ (***)
RandomLasso - Ridge	500	0.68 [0.66, 0.69]	91.60 ± 2.57	5213.59 ± 880.28	$<2.2*10^{-16}$ (***)
RandomLasso - Ridge	1000	0.68 [0.66, 0.70]	91.88 ± 2.59	10427.18 ± 1760.57	$<2.2*10^{-16}$ (***)
RandomLasso - Elnet	5	0.64 [0.60, 0.66]	90.68 ± 10.86	52.13 ± 8.80	0.39
RandomLasso - Elnet	10	0.64 [0.60, 0.67]	91.20 ± 7.57	104.27 ± 17.61	0.65
RandomLasso - Elnet	50	0.64 [0.60, 0.67]	91.06 ± 3.49	521.34 ± 88.03	0.44
RandomLasso - Elnet	100	0.64 [0.61, 0.67]	91.47 ± 3.08	1042.68 ± 176.06	$1.85*10^{-3}$ (**)
RandomLasso - Elnet	500	0.65 [0.61, 0.68]	91.60 ± 2.57	5213.41 ± 880.28	$3.83*10^{-5}$ (***)
RandomLasso - Elnet	1000	0.65 [0.62, 0.68]	91.88 ± 2.59	10426.82 ± 1760.55	$8.71*10^{-9}$ (***)

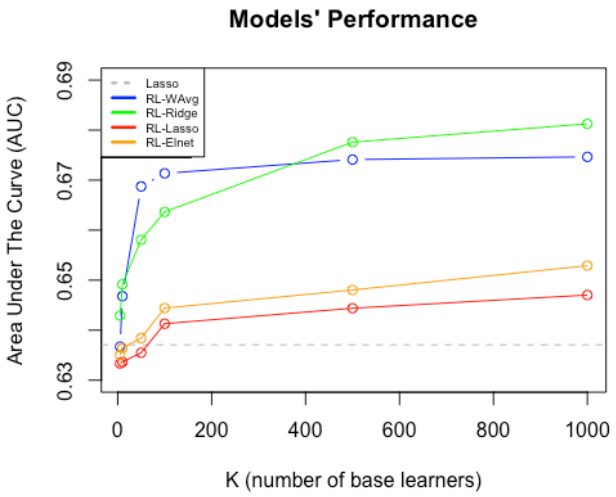

Fig. 3. Comparison of the AUC performance between the baseline LASSO model and each of the 4 ensemble approaches proposed with respect to the number of K base learners used.

represents the average number of genes retained across the K base learners, across folds and across repetitions of the validation strategy. Therefore, it can be logically argued that this measure of the ensemble approaches is not straightforward comparable to the number of genes retained by the standard LASSO. Despite this side note, authors considered worth noting the small number of genes retained in average across the base learner in order to achieve such a good improvement through the ensemble approach. Although it is not the scope of this paper, authors would like to point out that further work could be done within this line to analyze the gene frequency across base learners and perform a biological analysis of the genes with highest occurrences.

Another important factor to take into account when fitting ensembles is the execution time. As ensembles rely on the principle of fitting several base learners to data, it is easily deducted that these kind of models will demand much computing resources. This intuition is reflected in Table I where ensemble approaches require approximately a number of minutes approximately equals to $0.81 \times K \times T_{LASSO}$, where T_{LASSO} is the time required by the LASSO baseline model

(12.81 minutes).

V. Conclusions

This work has presented an ensemble approach which uses LASSO as base learners to analyze high dimensional data. Particularly, this paper applied the proposed approach to analyze a biomedical dataset of RNA-Seq gene expression profiles and compare its predictive performance to the one obtained with a standard LASSO model. More precisely, four different ensembles were proposed. The simplest one computed a weighted average of the individual predictions provided by each base learner according to their individual predictive performance. The remaining three ensemble methods consisted of meta-models based on linear models with some kind of regularization (l_1 only, l_2 only, or l_1 and l_2 combined) that learned the optimal combination of individual base learners' predictions.

The results of the analysis showed that any of the ensemble approaches proposed helped to outperform the standard LASSO in terms of AUC. In the worst case scenario, some configuration of the ensembles obtained an AUC of 0.65 compared to 0.64 of the standard LASSO. On the other hand, there were a few configurations of the ensemble that pushed the AUC up to 0.68, i.e. 0.04 points more than the baseline model. Furthermore, the performance of most of the ensemble approaches considered were statistically significant compared to the performance obtained by a standard LASSO according to the Wilcoxon signed rank test. An a priori disadvantage of ensemble methods is their demand of computing resources, something that can be easily overcome estimating the ensembles under a high performance cluster in a few days. Regarding which specific model to choose, we can extract from the obtained results (see Table I and Fig. 3) the following conclusions: if there is no strong computation time limitation (i.e., any K can be used) the best method seems to be $RandomLasso - Ridge$ (overall better results for K larger than 400), but if computation resources are moderate and a value lower than $K = 100$ has to be considered, then it seems reasonable to use the weighted average method as its performance grows steadily from $K = 1$ up to $K = 100$, outperforming the other models in this region.

The authors consider that this work could be potentially extended. Although Section III described an honest validation scheme to test the performance of the ensembles in future data, a further extension could consider to estimate the meta-model in the out-of-bag samples of an internal cross-validation procedure rather than doing it on the holdout dataset of the used validation strategy. This procedure should augment the robustness of the ensemble, thus obtaining more precise predictions and more confident AUCs. Moreover, it would be interesting to see whether the positive findings of this work replicate or not in other high dimensional datasets. Finally, richer models and other ensembles could be considered as possible base learners and meta-models to combine their individual predictions.

Acknowledgment

The authors acknowledge support through grants TIN2014-58516-C2-1-R from MICINN-SPAIN which include FEDER funds, and from ICE Andalucía TECH (Spain) through a postdoctoral fellowship. The authors also acknowledge support from the Universidad de Málaga ("Ayudas Plan Propio").

References

[1] K. R. Foster, R. Koprowski, and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research - commentary," *BioMedical Engineering OnLine*, vol. 13, no. 1, p. 94, 2014.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[3] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting breast cancer recurrence using machine learning techniques: A systematic review," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 52:1–52:40, 2016.

[4] P. Pinsky and C. Zhu, "Building multi-marker algorithms for disease prediction—the role of correlations among markers," *Biomarker Insights*, vol. 6, pp. 83–93, 2011.

[5] D. Urda, F. Aragón, L. Franco, F. J. Veredas, and J. M. Jerez, "L_1-regularization model enriched with biological knowledge," in *Bioinformatics and Biomedical Engineering - 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26-28, 2017, Proceedings, Part I*, 2017, pp. 579–590.

[6] C. Cui and D. Wang, "High dimensional data regression using lasso model and neural networks with random weights," *Inf. Sci.*, vol. 372, pp. 505–517, 2016.

[7] S. W. Purnami, S. Andari, and Y. D. Pertiwi, "High-dimensional data classification based on smooth support vector machines," *Procedia Computer Science*, vol. 72, pp. 477 – 484, 2015.

[8] X. Wang, "A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality," 2011, pp. 1293–1299.

[9] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse linear discriminant analysis by thresholding for high dimensional data," *Ann. Statist.*, vol. 39, no. 2, pp. 1241–1265, 2011.

[10] T. G. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*. London, UK, UK: Springer-Verlag, 2000, pp. 1–15.

[11] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *In Proceedings of the 21st International Conference on Machine Learning*. ACM Press, 2004, pp. 137–144.

[12] T.-N. Do, P. Lenca, S. Lallich, and N.-K. Pham, *Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees*, 2010, pp. 39–55.

[13] L. Song, P. Langfelder, and S. Horvath, "Random generalized linear model: a highly accurate and interpretable ensemble predictor," *BMC Bioinformatics*, vol. 14, no. 1, 2013.

[14] H. Wang, Q. Xu, and L. Zhou, "Large unbalanced credit scoring using lasso-logistic regression ensemble," *PLOS ONE*, vol. 10, no. 2, pp. 1–20, 2015.

[15] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[16] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 942–956, 2005.

[18] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome"," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 58, no. 1, pp. 267–288, 1996.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[21] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, 2004.
[22] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.
[23] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[24] A. Lacoste, F. Laviolette, and M. Marchand, "Bayesian comparison of machine learning algorithms on single and multiple datasets," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, vol. 22, 2012, pp. 665–675.