The Smallest Sample Size for the Desired Diagnosis Accuracy

ALEXANDRU G. FLOARES
SAIA & Artificial Intelligence Expert
Department of Biomedical Informatics
Vlahuta, Bloc Lama C, 45, Cluj
ROMANIA
alexandru.floares@saia-institute.org

MARIUS FERISGAN
SAIA & Artificial Intelligence Expert
Department of Biomedical Informatics
Vlahuta, Bloc Lama C, 45, Cluj
ROMANIA
marius.fersigan@saia-institute.org

DANIELA ONITA
SAIA & Artificial Intelligence Expert
Department of Biomedical Informatics
Vlahuta, Bloc Lama C, 45, Cluj
ROMANIA
danielaonita25@gmail.com

ANDREI CIUPARU
SAIA & Artificial Intelligence Expert
Department of Biomedical Informatics
Vlahuta, Bloc Lama C, 45, Cluj
ROMANIA
andrei.ciuparu@gmail.com

GEORGE A. CALIN
University of Texas, MD Anderson Cancer Center
Department of Experimental Therapeutics
Houston, TX
USA
gcalin@mdanderson.org

FLORIN B. MANOLACHE
Carnegie Mellon University
Department of Scientific Computing
Pittsburgh, PA
USA
florin@andrew.cmu.edu

Abstract: High-quality omics tests can be developed by using machine learning. As high-throughput molecular determinations are costly, we want to build the best models, utilizing the minimal number of samples. Here, we specify a set of criteria for high-quality models and select the algorithms which best satisfy them. Boosted C5, Random Forest and Stochastic Gradient Boosting reach accuracy greater than 95%, and even greater than 99%, in discriminating between breast cancer and normal, on the miRNA NGS TCGA data, generalize well to new cases, and are relatively transparent. For these algorithms, we investigate the relationships between accuracy and sample size, and between the number of features (miRNAs here) and sample size. We proposed power law formulas for all these relationships, allowing the computation of the required number of samples for the desired accuracy. The above algorithms dramatically lower the sample size for the highest accuracies and reduce the corresponding costs.

Key–Words: cancer, microRNA, next generation sequencing, machine learning, predictive models, power law

1 Introduction

Precision medicine cannot be reached just by increasing measurements' accuracy. We need adequate bioinformatics tools and workflows to analyze the data. These should include machine learning (ML) methods, as ML models can be translated in high-quality molecular tests for diagnosis, prognosis, and response to treatment prediction. For this goal, ML models should satisfy what we called ART Criteria: accuracy, robustness, and transparency. For most biomedical problems, as we show below, a 90%, 95% or even 99% accuracy is possible and desired. By robustness, we mean the capability of the predictive models to generalize well to new cases. Transparent, easy to understand, interpretable models, are also desired for molecular tests.

Thus, we have a multi-criteria optimization problem, where some criteria conflict each other. For example, a higher performance can be reached using ensemble methods (see, for example, [7] and [21]), but ensemble models are less interpretable. As cost is an important criterion too, we add it to the optimization problem. A proxy for cost is the number of samples. This is why literature is dominated by small studies, with an average of twenty samples for microarray measurements and six for Next Generation Sequencing (NGS).

Considering that the number of variables could be about two thousand for microRNA (miRNA) and

twenty thousand for mRNA, a sample size of six is too small for an ML approach. However, the goal of most omics studies is to find the differentially expressed genes between two or more biomedical situations. Even for such a modest goal, recent investigations [20], [4], showed that small sample size lead to unstable lists of differentially expression genes. While lists of differentially expressed genes have almost no translational impact, high-quality predictive models could have.

Thus, it is important to know how many samples we need for a 90%, 95% or even 99% accurate model, with different algorithms. It is also important to know the number of predictors for the specified accuracies. As both the sample size and the number of predictors are related to measurements' cost, we are interested in minimizing both, while maximizing the accuracy.

Another concern, which we are not addressing here, is related to the size of NGS data. For example, a FASTQ file is about 100 GB. Big Data pose transfer and storage problems. Finding the minimum number of samples for the desired accuracy alleviates these problems too.

Most studies investigate the relationships between differentially expressed genes, sample size and statistical power [13]-[15], [20], for various microarray datasets. In [20], similar relationships were investigated for predictive genes, discovered from data with LASSO, a classification algorithm with built-in feature selection. Support vector machine (SVM) was used in [17], and an explicit inverse power law formula was fitted, to model the relationship between SVM error and sample size. All these studies use relatively small microarray datasets.

In a recent paper [10], we used The Cancer Genome Atlas (TCGA) NGS miRNA determinations for breast cancer (generated by the TCGA Research Network: http://cancergenome.nih.gov/) to investigate these relationships. To the best of our knowledge, this is the largest dataset of this kind (865 sample). We used single powerful decision trees algorithms—C5 [18] and CART [1].

We found that the relationship between performance and sample size, and between the number of relevant predictors and sample size, can be modeled by power law formulas. Both C5 and CART have an accuracy greater than 95% for a samples size of about 100 patients. This is more than five-time greater than the average samples size. For smaller samples size, the accuracy is lower, but the models generalize well to new cases—C5 and CART models are robust. C5 benefits more than CART from samples size increase. The miRNA content of the discovered predictive genes set has higher stability when the samples size increases.

Here, we want to see if ensemble of decision trees—C5 with boosting [19], Random Forest [2], and Xgboost [3]—can reduce significantly the number of samples, for the desired performance, compared with single decision trees [10]. We also investigated the relationship between the sample size and the number of relevant miRNAs. All these relationships were accurately modeled with power law formulas. This allows easy and rational sample size planning, which is an important aspect of the design of experiments and could save important amounts of money.

2 Materials and Methods

Here we describe the dataset used, the preprocessing steps, the classification problem and the algorithms.

2.1 Datasets and Preprocessing

We used a subset of TCGA dataset, containing miRNA determinations, from normal and breast cancer tissue, totalizing 865 samples. We queried the normalized data from the GDC database [12], using the R package *TCGABioLinks*. We made use of the *caret* R package [16] for preprocessing steps. One of the main advantages of decision trees is that they need almost no preprocessing.

To eliminate irrelevant features and reduce the algorithms' learning time, we removed all variables with zero and near zero variance, using the *nearZeroVar* function from *caret* R package.

Even if our data suffered from a mild unbalancing (549 breast cancer samples and 316 normal tissue samples), we chose a stratified sampling strategy when partitioned data in training and testing sets. The reason for this is that we wanted to train our algorithms on datasets similar to those met in most study designs.

2.2 Classification Problem and Algorithms

We used the following ensemble of decision tree algorithms, which are powerful and meet the ART criteria: Xgboost, Random Forest and Boosted C5. To evaluate how the training sample size is related to the predictive performance, we used a repeated incremental stratified sampling method, starting from a sample size of 20 to a maximum training sample size of 600, by a step equal with 5. All the rest of the samples—validation set—were used for testing the generalization capability of the predictive models. For example, the smallest sample size is 20, and we tested the models developed on these samples on the remaining $865 - 20 = 845$ samples. Thus, we have the cross-validation (CV) performance and the performance on

all remaining samples. We focused mainly on the last one as a more reliable estimation for both the overfitting tendency and the generalization capability. This performance is included in all proposed formulas.

We trained the algorithms, on every training sample, maximizing the Receiver Operating Characteristic (ROC) [8] Area Under the Curve (AUC) metric. The Accuracy (ACC) is not optimized, but just computed for the corresponding ROC AUC, as the number of correct predictions from all predictions made. To estimate the out-of-sample performance, we made use of 3-fold CV. We used 3-fold CV instead of 10-fold CV because at small training sample sizes (e.g., 20, 25, 30) the condition "at least one different sample per class in every fold" didn't hold for 10-fold CV. For every training sample size, we performed 100 runs for every algorithm, averaging the metrics of interest (ROC AUC, number of selected features, Accuracy) for both CV and validation dataset.

There are two main advanced machine learning approaches for performance increasing: ensemble methods and hyperparameter optimization (optimizing algorithms' parameters). Here we focused mainly on the first, because the second cannot be performed together with learning curves. If the sample size range is large (here, from 20 to 600), the best parameter for small sample sizes will not be the same for medium and large sample sizes.

The main problem with small sample size is the high overfitting risk. Thus, the optimization will target the parameters capable of preventing overfitting, and we will favor those values reducing the risk of overfitting. For example, we will prefer trees with small depth. While this could prevent overfitting, it could also lead to underfitting for larger sample sizes. Thus, hyperparameter tuning could be performed either for each sample size or properly chosen sample size intervals. The first approach is computationally intensive and seems unjustified, as the optimal parameters do not change when the sample size is increased by just five. Our work on finding the proper intervals for parameter tuning is in progress.

3 Results and Discussions

Here, we analyze the results of using different ensemble decision trees algorithms—Boosted C5, Random Forest and Xgboost—on increasing sample sizes.

3.1 Performance - Sample Size Relationship

We are mainly interest in:

1. The relationships between algorithms' performances, measured as ROC AUC and Accuracy, and the sample size.
2. The relationships between the number of relevant features (miRNAs) and the sample size.

As we mentioned, ROC AUC was the objective function maximized by the three algorithms, not the Accuracy. This is because in bioinformatics ROC AUC is preferred for binary classification. Accuracy is just computed for the corresponding ROC AUC.

Table 1 presents a simple summary statistics of the three algorithms' performance, where we used the following abbreviations: RF for Random Forest, XGB for Xgboost, BC5 for Boosted C5, AUC for ROC AUC and ACC for Accuracy. Maximum ROC AUC and Accuracy are the same for all three algorithms. The mean is the same for RF and XGB and close for BC5. The minimum performance ranks the algorithms in the following order: RF, XGB, and BC5. Globally, RF looks like the most performant algorithm. However, we have to consider other criteria too:

1. The number of features for a given performance, especially for the high ones.
2. The difference between the cross-validation and validation performance.
3. The variability of the relevant miRNAs list of successive models from increasing sample size.

The number of features for increasing sample size will be analyzed in more details below. For small sample sizes, BC5 has smaller but reasonably good performances. However, the CV performance is almost identical with the validation one. This indicates that BC5 has no overfitting tendency. RF and XGB has greater performances for small sample size, but also a slight overfitting tendency.

The above-mentioned variability in miRNAs lists is decreasing with the sample size for all three algorithms (results not shown). This variability is very intriguing for the biomedical community. While this subject is outside the scope of this paper, it is important to mention that the main cause of this variability is the functional redundancy of the miRNome [11]. Thus, we do not have to worry about this variability. Most probably, the variability is inside the biologically relevant miRNAs' list.

From Figures 1 to 6 we can see that both AUC ROC and accuracy increase with sample size for all algorithms. The increase is higher for small data sizes and slows down for bigger data sizes.

We also analyzed the number of relevant features selected by every algorithm and its dependency on training sample size. From Figures 7 to 9 we can

Performance	Min	Mean	Max
RF AUC	0.92	0.97	0.99
RF ACC	0.94	0.97	0.99
XGB AUC	0.88	0.97	0.99
XGB ACC	0.90	0.97	0.99
BC5 AUC	0.81	0.96	0.99
BC5 ACC	0.83	0.96	0.99

Table 1: Algorithms Performance Statistics

No Features	Min	Mean	Max
RF	94	324	416
XGB	27	110	170
BC5	1	17	32

Table 2: Algorithms Relevant Features

see that the number of relevant features increases with sample size for all three algorithms. Boosted C5 selected 17 relevant features in average, with a minimum of 1 and maximum of 32. Random Forest selected an average of 324, with a minimum of 94 and maximum 416 relevant features. Xgboost used an average of 110 relevant features, with a minimum of 27 and maximum of 170, as represented in figure 10.

To obtain compact formulas for the performance and the number of relevant features dependency on the sample size, we tested most of the methods included in MATLAB Curve Fitting Toolbox (MATLAB R2017a and Curve Fitting Toolbox, The MathWorks, Inc., Natick, Massachusetts, United States). The best fit was a power law (the *power* option of the toolbox) formula for all algorithms and performance measurements. It fits a function of the form $y = ax^b + c$, where y represents the performance (ROC AUC or Accuracy), x represents the sample size, and a, b, and c are constant coefficients.

The power law formulas for the three algorithms will be presented in the following subsections.

3.2 Random Forest Power Law Formulas

The power law formulas for Random Forest are:

$$AUC = -1.646 \cdot (ss)^{-1.106} + 0.9847 \quad (1)$$

$$ACC = -1.049 \cdot (ss)^{-1.010} + 0.9862 \quad (2)$$

$$NOF = 4347 \cdot (ss)^{0.02141} - 4568 \quad (3)$$

where AUC represents the ROC AUC, ACC represents the accuracy, NOF represents the number of relevant features, ss the sample size, and the a, b, and c coefficients have their mean values.

The 95% confidence bounds for the coefficients are:

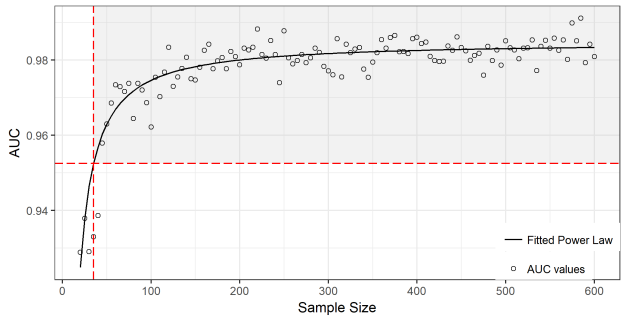

Figure 1: The Random Forest AUC versus the training sample size: results and fitted power law curve. The top-right quadrant delimited by red-dashed lines represents the area with computed AUC>0.95.

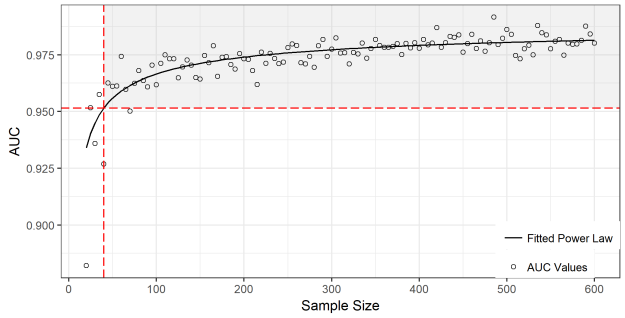

Figure 2: The Xgboost AUC versus the training sample size: results and fitted power law curve. The top-right quadrant delimited by red-dashed lines represents the area with computed AUC>0.95.

1. For equation 1: $a = (-1.939, -1.353)$, $b = (-1.16, -1.051)$, and $c = (0.9842, 0.9852)$
2. For equation 2: $a = (-1.215, -0.8833)$, $b = (-1.059, -0.9617)$, and $c = (0.9857, 0.9866)$
3. For equation 3: $a = (-7606, 1.63e + 04)$, $b = (-0.03169, 0.07451)$, and $c = (-1.659e + 04, 7450)$.

The goodness of fit tests' values are:

1. For equation 1: SSE : 0.0001562, $R-square$: 0.9888, $AdjustedR - square$: 0.9886, and $RMSE$: 0.00117.
2. For equation 2: SSE : 0.0001001, $R-square$: 0.9886, $AdjustedR - square$: 0.9884, and $RMSE$: 0.0009372.
3. For equation 3: SSE : 1.343e+04, $R-square$: 0.9825, $AdjustedR - square$: 0.9822, and $RMSE$: 10.85

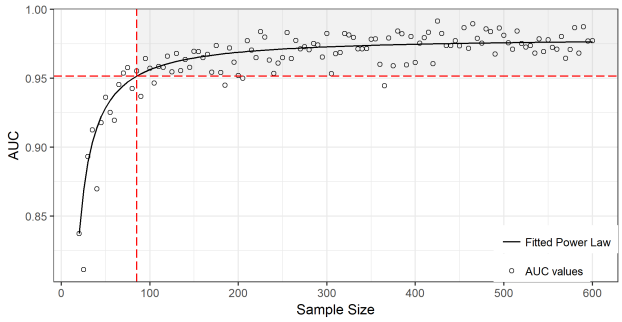

Figure 3: The BC5 AUC versus the training sample size: results and fitted power law curve. The top-right quadrant delimited by red-dashed lines represents the area with computed AUC>0.95.

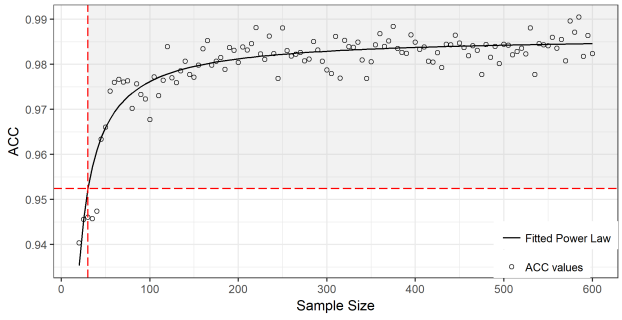

Figure 4: The Random Forest ACC versus the training sample size: results and fitted power law curve. The top-right quadrant delimited by red-dashed lines represents the area with computed ACC>0.95.

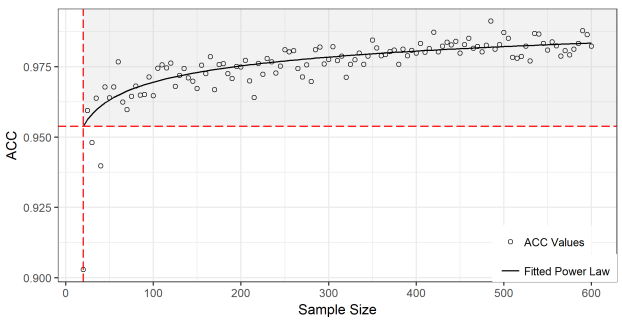

Figure 5: The Xgboost ACC versus the training sample size: results and fitted power law curve. The top-right quadrant delimited by red-dashed lines represents the area with computed ACC>0.95.

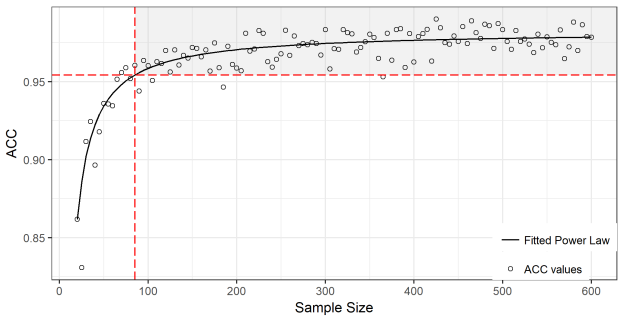

Figure 6: The BC5 ACC versus the training sample size: results and fitted power law curve. The top-right quadrant delimited by red-dashed lines represents the area with computed ACC>0.95.

3.3 Xgboost Power Law Formulas

The power law formulas for Xgboost are:

$$AUC = -0.2760 \cdot (ss)^{-0.5281} + 0.9908 \quad (4)$$

$$ACC = -0.1202 \cdot (ss)^{-0.1354} + 1.0340 \quad (5)$$

$$NOF = 5.524 \cdot (ss)^{0.5344} - 2.853 \quad (6)$$

The 95% confidence bounds for the coefficients are as follows:

1. For equation 4: $a = (-0.3296, -0.2224)$, $b = (-0.5984, -0.4577)$, and $c = (0.9879, 0.9936)$
2. For equation 5: $a = (-0.146, -0.0944)$, $b = (-0.2219, -0.04896)$, and $c = (0.9949, 1.073)$
3. For equation 6 are $a = (2.291, 8.757)$, $b = (0.4541, 0.6147)$, and $c = (-17.06, 11.35)$

The goodness of fit tests' values are:

1. For equation 4: SSE : 0.000332, $R-square$: 0.9821, $AdjustedR-square$: 0.9818, and $RMSE$: 0.001707.
2. For equation 5: SSE : 0.0002463, $R-square$: 0.9802, $AdjustedR-square$: 0.9798, and $RMSE$: 0.00147.
3. For equation 6: SSE : 4443, $R-square$: 0.9744, $AdjustedR-square$: 0.9739, and $RMSE$: 6.243.

3.4 Boosted C5 Power Law Formulas

The power law formulas for Boosted C5 are:

$$AUC = -4.1520 \cdot (ss)^{-1.1260} + 0.9795 \quad (7)$$

$$ACC = -2.4650 \cdot (ss)^{-1.007} + 0.9825 \quad (8)$$

$$NOF = 6.557 \cdot (ss)^{0.2995} - 17.16 \quad (9)$$

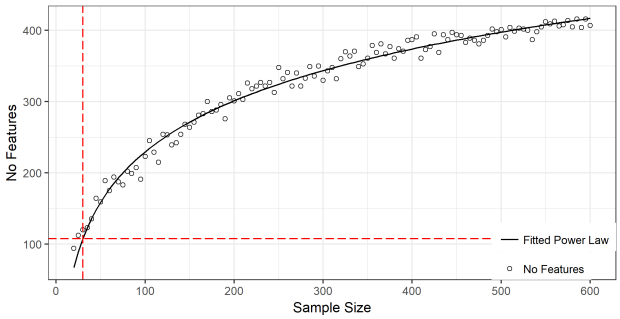

Figure 7: The Random Forest number of relevant features versus training sample size: results and fitted power law curve.

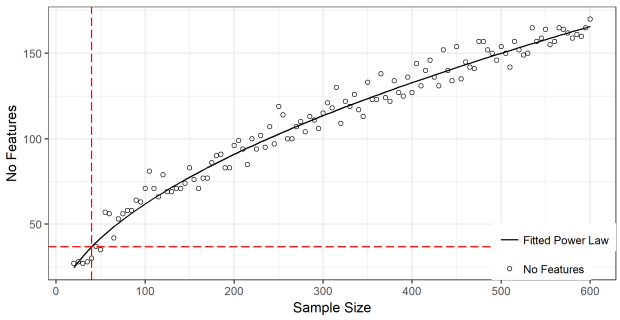

Figure 8: The Xgboost number of relevant features versus the training sample size: results and fitted power law curve.

The 95% confidence bounds for the coefficients are:

1. For equation 7: $a = (-4.924, -3.379)$, $b = (-1.183, -1.069)$, and $c = (0.9783, 0.9807)$
2. For equation 8: $a = (-8.472, 21.59)$, $b = (-1.062, -0.9521)$, and $c = (0.9812, 0.9837)$
3. For equation 9: $a = (-2.904, -2.026)$, $b = (0.03361, 0.5654)$, and $c = (-44.82, 10.5)$

The goodness of fit tests' values are:

1. For equation 7: SSE : 0.0009371, $R-square$: 0.9883, $AdjustedR-square$: 0.9881, and $RMSE$: 0.002867
2. For equation 8: SSE : 0.0007194, $R-square$: 0.9877, $AdjustedR-square$: 0.9875, and $RMSE$: 0.002512
3. For equation 9: SSE : 2166, $R-square$: 0.703, $AdjustedR-square$: 0.6978, and $RMSE$: 4.359

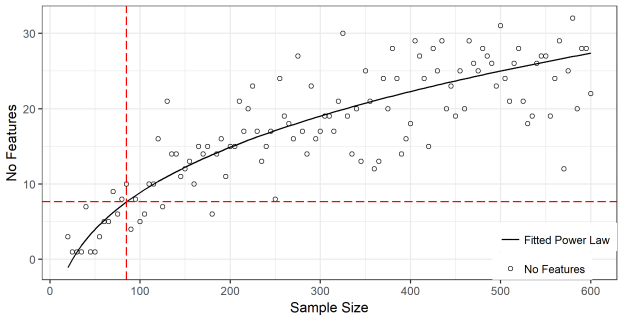

Figure 9: The BC5 ACC versus the training sample size: results and fitted power law curve.

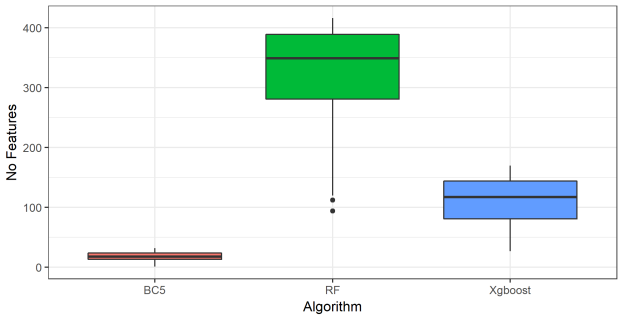

Figure 10: Distribution of relevant features for BC5, Random Forest and Xbgoost algorithm.

4 Conclusion

We found that using ensembles of decision trees, like Boosted C5, Random Forest, and Stochastic Gradient Boosting, we can develop high-quality models—accurate, robust (generalizing well to new cases), and relatively transparent—from omics data. This can be translated in diagnosis, prognosis, and response to treatment molecular tests. We also found that using these algorithms, we need the smallest sample size for the highest accuracies. This dramatic decrease in the required sample size is reflected in a significant cost decrease. We derived power law formulas which can be effectively used in design of omics experiments.

Acknowledgements: This work was supported by the research grants UEFISCDI PN-II-PT-PCCA-2013-4-1959 INTELCOR and UEFISCDI PN-II-PT-PCCA-2011-3.1-1221 IntelUro, financed by Romanian Ministry of Education and Scientific Research.

References:

[1] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees.

The Wadsworth statisticsprobability series 19, 1984

[2] L. Breiman, Random forests. *Machine Learning* 45(1), 2001, pp. 5-32. https://doi.org/10.1023/A:1010933404324

[3] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System. CoRR, abs/1603.0., 2016, Retrieved from http://arxiv.org/abs/1603.02754

[4] T. Ching, S. Huang, L.X. Garmire, Power analysis and sample size estimation for RNA-Seq differential expression, *RNA* 20, 11, 2014, pp. 1684–96

[5] A. Clauset, C. R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data. *SIAM Review* 51(4), 2009, pp. 661-703.

[6] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, H. Noushmehr, TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data, *Nucleic Acids Research*, 2015

[7] T. G. Dietterich, Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 2000

[8] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Researchers, 2004

[9] M. Fernndez-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* 15, 2014, pp. 3133-3181.

[10] A. G. Floares, G. A. Calin, F. B. Manolache, Bigger Data Is Better for Molecular Diagnosis Tests Based on Decision Trees BT - *Data Mining and Big Data: First International Conference, DMBD 2016, Bali, Indonesia, June 25-30, 2016. Proceedings*. In Y. Tan Y. Shi (Eds.) (pp. 288295). Cham: Springer International Publishing. (Conference Best Paper Award), 2016

[11] A. G. Floares, C. Braicu, R. Cojocneanu-Petric, I. B. Neagoe, G. A. Calin, L. Adam, F. Manolache, Exploring the Functional Redundancy of miRNA in Cancer with Computational Intelligence. *Proceedings of the Computational Intelligence in Bioinformatics and Biostaistics 2016*, Stirling, UK, 2016.

[12] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, L. M. Staudt, Toward a Shared Vision for Cancer Genomic Data, *New England Journal of Medicine* 375:12, 2016, pp. 1109–1112

[13] SH. Jung, Sample size for FDR-control in microarray data analysis, *Bioinformatics* 21, 14, 3097–3104, 2005

[14] SH. Jung, SS. Young, Power and sample size calculation for microarray studies, *J Biopharm Stat.* 22(1), 30–42, 2012

[15] SH. Jung, H. Bang, SS. Young, Sample size calculation for multiple testing in microarray data analysis, *Biostatistics* 6, 1, 2005, pp. 157-169

[16] M. Kuhn. Contributions from J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt, caret: Classification and Regression Training. R package version 6.0-76, 2017

[17] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, J. P. Mesirov, Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 10(2), 2003, 119142. https://doi.org/10.1089/106652703321825928

[18] J. R. Quinlan, C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., 1993

[19] R. E. Schapire, Y. Freund, Boosting: Foundations and Algorithms. *The MIT Press*, 2012

[20] - C. Stretch, S. Khan, N. Asgarian, R. Eisner, S. Vaisipour, S. Damaraju, et al,Effects of Sample Size on Differential Gene Expression, Rank Order and Prediction Accuracy of a Gene Signature, *PLoS ONE* 8(6), 2013

[21] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms,*, CRC Press 2012. https://doi.org/doi:10.1201/b12207-2