An Evolutionary Approach to the Discretization of Gene Expression Profiles to Predict the Severity of COVID-19

Nisrine Mouhrim[1], Alberto Tonda[1], Itzel Rodríguez-Guerra[2],
Aletta D. Kraneveld[3] and Alejandro Lopez Rincon[3],
[1]UMR 518 MIA-Paris, INRAE, Paris, France
[2]Laboratorio de bioquímica y genética microbiana del instituto de ciencias de la universidad autónoma de Puebla (ICUAP) Puebla, Mexico
[3]Division of Pharmacology, University of Utrecht Utrecht, The Netherland

ABSTRACT

In this work, we propose to use a state-of-the-art evolutionary algorithm to set the discretization thresholds for gene expression profiles, using feedback from a classifier in order to maximize the accuracy of the predictions based on the discretized gene expression levels, while at the same time minimizing the number of different profiles obtained, to ease the understanding of the expert. The methodology is applied to a dataset containing COVID-19 patients that developed either mild or severe symptoms. The results show that the evolutionary approach performs better than a traditional discretization based on statistical analysis, and that it does preserve the sense-making necessary for practitioners to trust the results.

CCS CONCEPTS

• **Applied computing → Life and medical sciences**; • **Computing methodologies → Machine learning**; **Bio-inspired approaches**.

KEYWORDS

discretization, gene expression profiles, evolutionary optimization, covid-19, prognosis

ACM Reference Format:
Nisrine Mouhrim[1], Alberto Tonda[1], Itzel Rodríguez-Guerra[2],, Aletta D. Kraneveld[3] and Alejandro Lopez Rincon[3],, [1]UMR 518 MIA-Paris, INRAE, Paris, France, [2]Laboratorio de bioquímica y genética microbiana del instituto de ciencias de la universidad autónoma de Puebla (ICUAP) Puebla, Mexico, [3]Division of Pharmacology, University of Utrecht Utrecht, The Netherland . 2022. An Evolutionary Approach to the Discretization of Gene Expression Profiles to Predict the Severity of COVID-19. In *Genetic and Evolutionary Computation Conference Companion (GECCO '22 Companion), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3520304.3529001

1 INTRODUCTION

As the SARS-CoV-2 pandemic continues, there have been 378 million cases with 5.67 million deaths as of February 1st, 2022 [11]. Given the size of the outbreak, one of the major problems is the lack of the necessary medical equipment for care of patients, in contrast to the number of infected people. Considering the number of cases, different studies have tried to elucidate the differences between severe and mild cases using omics data, to quickly predict whether a patient will be in need of intensive care, and efficiently allocate the available medical resources: literature reports several examples, using DNA methylation [5], mRNA gene expression and/or microRNA data. The correct allocation and availability of hospital beds is considered a crucial factor for lowering COVID-19 mortality rates by several sources in literature [3].

Although the use of multi-omics data had had different degrees of success for diagnostic and prognostic purposes in general, translating the results into meaningful diagnostic tests or biomarkers for clinical practice is still challenging. To make sense of the data, medical practitioners often resort to the creation of *gene expression profiles*, discretizations of gene expressions, where each value is assigned to a category, for example under- or over-expressed. Categories are usually evaluated using thresholds based on the mean values of gene expressions from healthy controls as a baseline. While this procedure can help the sense making of the experts, such a discretization leads to loss of information and could potentially impair classification performance.

In this paper, we propose an evolutionary approach to the discretization of gene expression data, in order to obtain gene expression profiles that both provide good classification results and can be easily interpreted by domain experts. To this aim, we set a state-of-the-art evolutionary algorithm to optimize the threshold for discretization of each gene expression, aiming to maximize classification accuracy after discretization, and at the same time minimizing the number of different patient profiles produced.

The proposed approach is tested on real-world data from 138 patients, including the information from 60,671 genes, and compared against a more classical discretization approach based on mean values of gene expression from healthy controls. The results show that the methodology is effective in identifying 12 genes that are highly correlated with responsiveness to treatment, and it is able to discretize their gene expression levels into gene expression profiles, not only helping to increase the classification accuracy, but offering a human-interpretable explanation of the development of mild or dire symptoms from a COVID-19 infection.

2 PROPOSED APPROACH

We introduce a novel approach for the discretization of gene expression data to obtain gene expression profiles, interpretable by

GECCO '22 Companion, July 9–13, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9268-6/22/07.
https://doi.org/10.1145/3520304.3529001

domain experts. After a step of feature selection, the most relevant genes are discretized with thresholds optimized by an EA, with the objectives of maximizing classification accuracy and minimizing the number of different profiles in the discretized dataset.

2.1 Feature Selection

In a first step, our objective is to select the most meaningful genes to correctly predict and model COVID-19 patients' severity (mild/severe). We apply the REFS algorithm, which uses the feedback of an ensemble of classifiers to rank each feature depending on its usefulness for the process of classification. Then, the lowest-scoring features are removed, and the classification/ranking is repeated, until the average classification accuracy falls below a user-defined threshold. REFS is usually run multiple times, due to the random values used by some of the classifiers included in the ensemble.

2.2 Evolutionary Discretization

From the results of REFS algorithms we obtain a set of features. Nevertheless the values of the selected features are difficult to read for a clinician to take decisions. Therefore, instead of showing each feature as a continuous value, we use EAs to categorize the values into under and over expression, optimizing the thresholds for each selected feature (gene).

Once we have a reduced set of variables $V = \{v_0, v_1, v_2, ..., v_n\}$, given by the REFS algorithm, we use EAs to transform the input variables into *over* and *under* expressed values, labeled as 0 and 1, respectively: in other words, the EA will generate a vector of thresholds $I = \{t_0, t_1, t_2, ..., t_n\}$ to discretize each variable. A visual example of the transformation from a gene expression dataset to gene expression profiles is reported in Fig. 1.

The discretization will be optimized with respect to two criteria: classification accuracy, to be maximized; and number of different profiles, to be minimized. Thus, the fitness function for an individual I will be given by:

$$f(I) = w_p \cdot \frac{1.0}{1.0 + A(X, y)} + w_r \cdot n_p \qquad (1)$$

where X is the discretized dataset, y is the vector with the known labels (mild/severe symptoms) for each sample, $A(X, y)$ is the classification accuracy (number of correct label predictions by a classifier over total number of samples), n_p is the number of different profiles in the dataset after discretization, and w_p and w_r are weights. The fitness function is to be minimized, as the ideal candidate solution features both high accuracy and low number of different profiles, to ease the sensemaking of the domain experts.

3 EXPERIMENTAL EVALUATION

The proposed approach is implemented in Python 3, relying on the `cma` package for CMA-ES and the `scikit-learn` [9] package for classification. All the code and data needed to reproduce the experiments is freely available in the GitHub repository: https://github.com/albertotonda/ea-discretization-health.

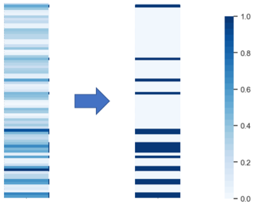

Figure 1: Example of discretization of gene expression data to gene expression profiles, using a threshold value of 0.46 to distinguish between over- and under-expressed genes.

3.1 Data

From the gene expression omnibus (GEO) repository, we select dataset GSE169687 that contains 138 samples of mRNA from peripheral blood of COVID-19 patients at different time points, and 14 healthy controls. We only consider the 138 samples from patients, with either mild/moderate (n=109) or severe/critical (n=29) symptoms. For each sample we have 60,671 ensemble genes, and we divide the dataset into 2 groups, where we assign the label 0 to patients with mild/moderate symptoms and 1 to patients with severe/critical symptoms.

3.2 Feature Selection

We run the REFS algorithm 10 times, and we get a reduction from 60,671 to 12 features (highlighted as the optimal trade-off in Fig. 2). This translates to the expression levels shown in Fig. 3.

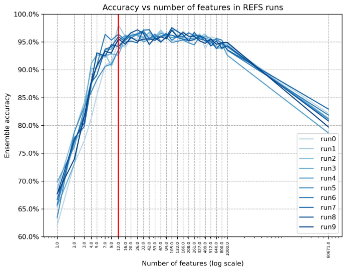

Figure 2: 10 runs of the REFS algorithm. The solution considered as the best compromise between accuracy and number of features is marked with a red line (n=12).

3.3 Profile Generator

The EA selected for the profile generator is the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [7], considered the state of the art for numerical optimization of non-convex functions with continuous values. After a few trial runs, the algorithm is set with the following parameters: $N = 12$, $\mu = 100$, $x_0 = \{0.5, 0.5, ...0.5\}$, $\sigma_0 = 0.1$, $w_p = 1.0$, $w_r = 10^{-5}$, all default stop conditions, and Logistic Regression as the classifier chosen to compute classification accuracy A for the fitness function described in Eq. 1. The choice of Logistic Regression is motivated by its effectiveness and training

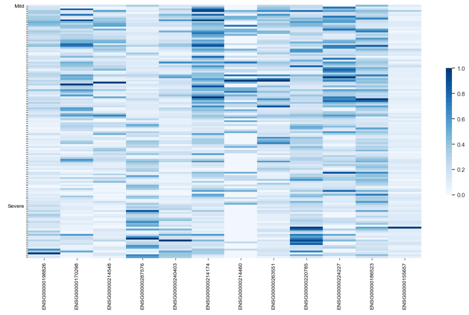

Figure 3: Heatmap of normalized gene expression data, showing the values of each patient for the 12 most important features selected by the REFS algorithm.

speed, making it one of the most suitable algorithms for our scenario. The classifier is run in a 10-fold cross-validation at each evaluation, in order to obtain a more reliable estimate of accuracy.

In order to provide a comparison, we also compute profiles based on a classical technique of the domain, using the gene expression levels of the healthy controls as a reference to discretize the gene expressions of the patients.

A possible downside to our methodology is to create discretization thresholds that are uniquely tailored to the classifier used for the fitness function, in this case, Logistic Regression, and lose generality. In order to test the generality of our method, we transform the data using all the resulting thresholds and compute mean accuracy in a 10-fold cross-validation, using Logistic Regression and seven other state-of-the-art classifiers. The results available in Table 1 show that even classifiers not included in the fitness function show high levels of accuracy, providing evidence against overfitting. The number of profiles obtained by each discretization, in comparison to the global accuracy is shown in Fig. 4.

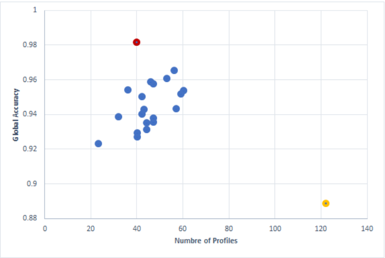

Figure 4: Runs of the EA profile generator. The best solution found by the proposed approach is marked in red (n=48). The solution computed using the classical technique of considering the gene expression levels of healthy controls as references is marked in orange (n=122).

From the results, we consider the best run as the solution generating 48 different profiles, with an average classification accuracy of 0.9817 over the 8 classifiers. Our framework clearly outperforms the more traditional approach of discretizing profiles based uniquely on the gene expression data from healthy controls, both from the point of view of classification accuracy and number of different profiles created after discretization. Transforming the dataset using the thresholds of the best individual found during the 20 experimental runs (Table 2) results in the heatmap shown in Fig. 5 where we have only values of 0 or 1, corresponding to under- or over-expressed genes.

Table 1: Discretization strategies compared, using the accuracy from different state-of-the-art classifiers, in a 10-fold cross-validation. Mean and (stdev) indicate the mean results obtained by the best individual of each of the 20 runs of the EA profile generator. Healthy indicates the discretization performed using healthy controls as reference for the thresholds in the gene expression levels. Best is the performance of the best individual produced by the EA profile generator. 9 Genes Transformation shows again the performance of the best individual produced by the EA profile generator, considering only 9 genes instead of 12

	Mean	(stdev)	Healthy	Best	9 Genes Transformation
Gradient Boosting [6]	0.9395	(0.0200)	0.8846	0.9703	0.9484
Random Forest [2]	0.9461	(0.0171)	0.9137	0.9703	0.9637
Logistic Regression	0.9608	(0.0198)	0.9060	0.9929	0.9929
Passive Aggressive Classifie [4]	0.9403	(0.0291)	0.8484	0.9929	1.0000
Stochastic Gradient Descent [13]	0.9344	(0.0298)	0.8764	0.9780	0.9709
Support Vector Machines (linear) [10]	0.9768	(0.0147)	0.8978	1.0000	0.9929
Ridge [8]	0.9372	(0.0203)	0.8918	0.9786	0.9418
Bagging [1]	0.9344	(0.0179)	0.8923	0.9703	0.9484
Mean accuracy	0.9462	0.0143	0.8889	0.9817	0.9699
Mean #Profiles	45.4	8.9129	122.0	48.0	45.0

Table 2: Generated thresholds for each of selected gene, considering the best individual obtained over the 20 runs.

Ensemble ID	Gene ID	Thresholds
ENSG00000198826	ARHGAP11A	0.0032
ENSG00000170298	LGALS9B	0.3639
ENSG00000214548	MEG3	0.1874
ENSG00000287576	-	0.4734
ENSG00000240403	KIR3DL2	0.1129
ENSG00000214174	AMZ2P1	0.0005
ENSG00000214460	TPT1P6	0.0746
ENSG00000263551	-	0.2120
ENSG00000220785	MTMR9LP	0.6295
ENSG00000224227	OR2L1P	0.0371
ENSG00000186523	FAM86B1	0.0012
ENSG00000155657	TTN	0.3101

Now, using the transformed dataset with the best thresholds found (Table 2) we are able to apply the REFS algorithm again (Fig. 6), which reduces even further the number of genes to (n=9) to separate between the two groups with a global accuracy of 0.9699, reducing even further the number of profiles from 48 to 45 (Fig. 7).

4 DISCUSSION

The results presented in Table 1 clearly show that the proposed approach outperforms the more traditional discretization methodology, both concerning the classification accuracy and the number

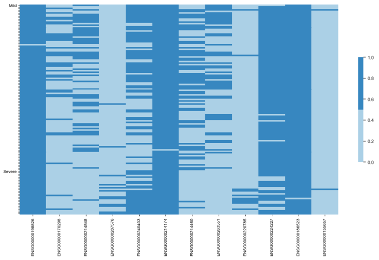

Figure 5: Heatmap produced by applying the set of thresholds identified by the best individual produced by the EA profile generator to the considered dataset.

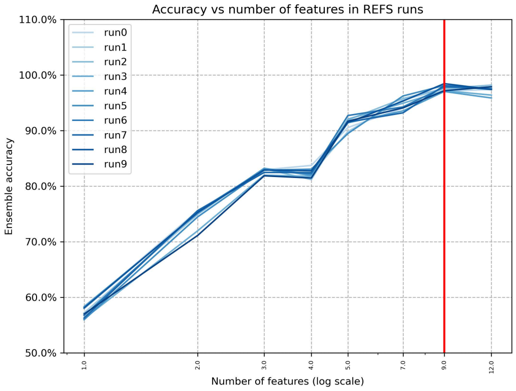

Figure 6: 10 runs of the REFS algorithm on the dataset discretized by the thresholds found by the best individual. The solution with the best compromise between accuracy and number of features is marked with the red line at n=9.

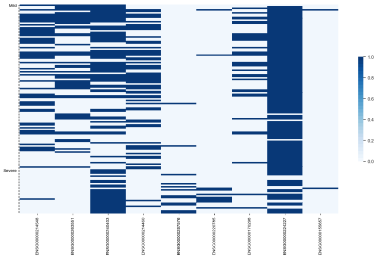

Figure 7: Heatmap obtained after applying the thresholds represented by the best individual found to the dataset, considering only the 9 most representative genes.

of profiles produced after discretization. Furthermore, the excellent accuracy results in a 10-fold cross-validation for several classifier provide evidence that using just Logistic Regression as part of the fitness function does not overfit the discretization thresholds to a single classifier.

Of the 12 genes selected by the proposed approach, 2 of them are novel transcripts: ENSG00000263551 and ENSG00000287576. Both are listed as lncRNA (long, non-coding RNA) in gene cards database [12], and there is no information available related to the subject matter, as for ENSG00000214460 (TPT1P6 gene). This could potentially be a lead for new research on the subject, as they have never been previously associated with any particular biological function in literature, to the best of our knowledge.

5 CONCLUSIONS AND FUTURE WORKS

In this paper, we presented a novel evolutionary approach to the discretization of gene expression data, in order to obtain interpretable gene expression profiles, that can also lead to good classification accuracy when used with ML classifiers. The results on a real-world dataset related to COVID-19, with patients exhibiting either mild or severe symptoms, seem promising, with the proposed technique performing better than a more classical approach based on a comparison with healthy controls. In addition, we generated a set of rules given 9 specific genes to be used as a guide to decide whether a patient will present severe symptoms.

While very promising, the results from 20 repeated runs of the proposed approach show some variance in both accuracy and number of different profiles obtained by the discretization of the gene expression dataset.

Future works will include experiments with a multi-objective fitness function, and testing the proposed methodology on different real-world applications in the health domain.

REFERENCES

[1] Leo Breiman. 1999. Pasting small votes for classification in large databases and on-line. *Machine Learning* 36, 1-2 (1999), 85–103.

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[3] Joseph J. Cavallo, Daniel A. Donoho, and Howard P. Forman. 2020. Hospital Capacity and Operations in the Coronavirus Disease 2019 (COVID-19) Pandemic—Planning for the Nth Patient. *JAMA Health Forum* 1, 3 (March 2020), e200345. https://doi.org/10.1001/jamahealthforum.2020.0345

[4] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7, Mar (2006), 551–585.

[5] Manuel Castro de Moura, Veronica Davalos, Laura Planas-Serra, Damiana Alvarez-Errico, Carles Arribas, Montserrat Ruiz, Sergio Aguilera-Albesa, Jesús Troya, Juan Valencia-Ramos, Valentina Vélez-Santamaria, et al. 2021. Epigenome-wide association study of COVID-19 severity with respiratory failure. *EBioMedicine* 66 (2021), 103339.

[6] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[7] N. Hansen and A. Ostermeier. 2001. Completely Derandomized Self-adaptation in Evolution Strategies. *Evolutionary computation* 9, 2 (2001), 159–195. https://doi.org/10.1063/1.2713540

[8] Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 1 (1970), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[10] John Platt and Others. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[11] Max Roser. 2022. Covid-19 data explorer. https://ourworldindata.org/explorers/coronavirus-data-explorer

[12] Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, et al. 2010. GeneCards Version 3: the human gene integrator. *Database* 2010 (2010).

[13] Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Twenty-first international conference on Machine learning - ICML '04*. ACM Press. https://doi.org/10.1145/1015330.1015332