POSTER

Batch correction of genomic data in chronic fatigue syndrome using CMA-ES

ALEJANDRO LOPEZ RINCON, Utrecht University, Utrecht, Netherlands

ALETTA D KRANEVELD, Utrecht University, Utrecht, Netherlands

ALBERTO PAOLO TONDA, National Research Institute for Agriculture, Food and Environment, Paris, Ile-de-France, France

Published: 08 July 2020

GECCO '20: Genetic and Evolutionary Computation Conference
July 8 - 12, 2020
Cancún, Mexico

Conference Sponsors:
SIGEVO

https://doi.org/10.1145/3377929.3389947
ISBN: 9781450371278

Batch Correction of Genomic Data in Chronic Fatigue Syndrome Using CMA-ES

Alejandro Lopez Rincon
Utrecht University
Utrecht, The Netherlands
a.lopezrincon@uu.nl

Aletta D. Kraneveld
Utrecht University
Utrecht, The Netherlands
A.D.Kraneveld@uu.nl

Alberto Tonda
UMR 518 MIA, INRAE
Paris, France
alberto.tonda@inrae.fr

ABSTRACT

Modern genomic sequencing machines can measure thousands of probes from different specimens. Nevertheless, theoretically comparable datasets can show considerably distinguishable properties, depending on both platform and specimen, a phenomenon known as *batch effect*. Batch correction is the technique aiming at removing this effect from the data. A possible approach to batch correction is to find a transformation function between different datasets, but optimizing the weights of such a function is not trivial: As there is no explicit gradient to follow, traditional optimization techniques would fail. In this work, we propose to use a state-of-the-art evolutionary algorithm, Covariance Matrix Adaptation Evolution Strategy, to optimize the weights of a transformation function for batch correction. The fitness function is driven by the classification accuracy of an ensemble of algorithms on the transformed data. The case study selected to test the proposed approach is mRNA gene expression data of Chronic Fatigue Syndrome, a disease for which there is currently no established diagnostic test. The transformation function obtained from three datasets, produced from different specimens, remarkably improves the performance of classifiers on the task of diagnosing Chronic Fatigue. The presented results are an important steppingstone towards a reliable diagnostic test for this syndrome.

ACM Reference Format:
Alejandro Lopez Rincon, Aletta D. Kraneveld, and Alberto Tonda. 2020. Batch Correction of Genomic Data in Chronic Fatigue Syndrome Using CMA-ES. In *Genetic and Evolutionary Computation Conference Companion (GECCO '20 Companion), July 8–12, 2020, Cancún, Mexico.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3377929.3389947

1 INTRODUCTION

Chronic Fatigue Syndrome (CFS) is a rare condition, with a worldwide prevalence of approximately 0.76 to 3.28% [11]. CFS is currently diagnosed based on symptoms, and there is no physical test able to detect the disease. Typically, CFS will be diagnosed after discarding all other possible causes of fatigue. Even as datasets on CFS are becoming more and more available to the bioinformatics community [4, 5, 16], thus opening the way to the development of an automated diagnostic test, one of the issues of genomic data collection is the heterogeneity of the results, making it difficult to compare between datasets that should be similar. In specialized literature, this issue is called *batch effect*, as biological datasets are also termed *batches*.

GECCO '20 Companion, July 8–12, 2020, Cancún, Mexico

ACM ISBN 978-1-4503-7127-8/20/07.
https://doi.org/10.1145/3377929.3389947

2 BACKGROUND

Despite the availability of data, using mRNA expression data in machine learning presents several challenges. First, the number of features against the number of samples; machines like Affymetrix Human Genome U133A Array have 22,283 probes to measure mRNA expressions. In contrast, the number of samples in CFS publicly available datasets is less than 1% of the number of features. In addition, mRNA studies show poor correlation between different measuring platforms and specimen type (PBMCs, PBLs, WB) [3, 12], implying a strong prevalence of the batch effect. To compensate for different specimen types and different measuring platforms, batch correction algorithms are thus necessary. Several batch correction algorithms are proposed in literature, with reviews comparing their relative effectiveness [2, 13]. In this work, we will describe in more detail *Combating Batch Effects When Combining Batches of Gene Expression Microarray Data* (ComBat) [10] and *Linear Models for Microarray Data* (Limma) [17]. ComBat uses an empirical Bayesian method to estimate the parameters of location and scale adjustment for each gene independently, with the advantage of not requiring a large amount of data samples to compute the batch correction. Limma, on the other hand, assumes that the batch effect can be represented by a linear model: It first fits the parameters of the batch effect on the available samples, and then removes the result from the data.

3 PROPOSED APPROACH

In the domain of genomics, it is accepted that the main difference between corresponding samples obtained from different cell types is mainly an issue of intensity of expression; that is, the transformation between two corresponding features in two different feature spaces mainly depends on a multiplicative factor. In defining the new problem:

$$S_B = T \times S_A = T' \times I \times S_A \tag{1}$$

where I is the identity matrix, and T' is a vector of size N. In other words, we assume that the transformation described by T can be limited to an anisotropic scaling, thus reducing the number of coefficients w to be set from N^2 to just N. This assumption might also potentially help in reducing overfitting issues, as often the amount of samples at hand might be less than N^2, and optimizing a transformation function with such a disproportionate number of parameters would greatly impair generalization.

The problem of finding the coefficients for T' does not have a clearly defined gradient, thus classical optimization techniques such as Broyden–Fletcher–Goldfarb–Shanno or stochastic gradient descent would fail. For this reason, we choose the state-of-the-art in gradient-free optimization, CMA-ES [8]. In this context, the **genotype** of a candidate solution is a vector of N real values. The **fitness function** to be maximized is the average accuracy of an ensemble of classifiers, in a k-fold stratified cross-validation over reference dataset B and the transformed samples of dataset A.

4 EXPERIMENTAL EVALUATION

Table 1 lists the datasets used in the experiments, containing gene expressions from patients either healthy or affected by Chronic Fatigue Syndrome. For each dataset, the data has been collected from a different specimen type. All the code for the experiments is available in a GitHub repository[1].

Table 1: Datasets used in the experimental evaluation.

Dataset [ref]	Specimen type	#samples	#healthy	#cfs
CAMDA2006 [15]	PBMC	118	25	93
GSE16059 [1]	PBL	88	44	44
GSE98139 [14]	Whole blood	47	18	29

4.1 Experimental results

The proposed approach and the two reference batch correction algorithms, ComBat and Limma, are run twice each, to transform datasets GSE16059 and GSE98139 in the feature space of dataset CAMDA2006. A numerical evaluation is then presented in Table 2, where CAMDA2006 and the two datasets transformed by each batch correction algorithm are evaluated in a stratified 10-fold cross-validation, with 4 classification algorithms not included in the ensemble of the proposed approach. The proposed approach shows a significant improvement over both the unaltered datasets, and the modifications proposed by ComBat and Limma, empirically showing how the transformation function T' learned by CMA-ES is better at placing samples in parts of the feature space corresponding to the two classes of the problem. Analyzing the Receiver Operating Characteristic (ROC) curves, we calculate the Area under the curve (AUC), which is a further measure of how good a classifier is at discriminating between CFS patients and healthy controls. Calculating the AUC for each of the 4 classifiers (AdaBoost, ExtraTrees, MLP and Perceptron) gives as result 0.8, 0.76, 0.98 and 0.98 respectively in a 10-fold cross validation. Following the guidelines, we have two classifiers with excellent diagnostic accuracy (AUC 0.9-1.0), one with very good (AUC 0.8-0.9), and one with good accuracy (AUC 0.7-0.8) [18]. Observing the time needed for the experiments, computational effort is the clear disadvantage of the proposed technique (>48h) with respect to ComBat (0.57s) and Limma (9.19s): The ensemble evaluation of a candidate solution requires several iterations of training and test using all the selected classifiers. It must be noted, however, that the current prototype implementation runs in a single thread, and a considerable speedup might be possible by parallelizing the evaluations at each generation of CMA-ES.

[1] https://github.com/albertotonda/ea-cluster-alejandro

Table 2: Classification results of the datasets after the batch correction techniques, compared to the initial situation, with no correction. Mean values marked with * are statistically better than the others, using a Welch's t-test ($p < 0.05$).

Classifier	No correction		ComBat		Limma		Proposed approach	
	mean	stdev	mean	stdev	mean	stdev	mean	stdev
AdaBoost [19]	0.7624	0.0700	0.7990	0.0867	0.8422	0.0523	0.8663*	0.0726
ExtraTrees [7]	0.7868	0.0789	0.7783	0.0289	0.7718	0.0757	0.8217*	0.0571
MLP [9]	0.7077	0.0522	0.9124	0.0472	0.7205	0.0571	0.9920*	0.0160
Perceptron [6]	0.7160	0.0713	0.9199	0.0702	0.6644	0.0646	0.9604*	0.0302
Mean	0.7432		0.8524		0.7497		0.9101	

5 CONCLUSIONS

The experimental results show that the proposed approach can obtain improvements in classification accuracy superior to two other state-of-the-art batch correction techniques, with several classifiers crossing the threshold of $AUC > 0.90$, often considered as *oustanding* for real-world applications in the health sector.

REFERENCES

[1] Andrea Byrnes et al. 2009. Gene expression in peripheral blood leukocytes in monozygotic twins discordant for chronic fatigue: no evidence of a biomarker. *PLoS One* 4, 6 (2009).

[2] Chao Chen et al. 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one* 6, 2 (2011).

[3] Duncan E Donohue et al. 2019. Gene expression profiling of whole blood: A comparative assessment of RNA-stabilizing collection methods. *PloS one* 14, 10 (2019).

[4] Ron Edgar, Michael Domrachev, and Alex E Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 30, 1 (2002), 207–210.

[5] Hong Fang et al. 2006. Gene expression profile exploration of a large dataset on chronic fatigue syndrome. (2006).

[6] Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning* 37, 3 (1999), 277–296.

[7] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning* 63, 1 (2006), 3–42.

[8] 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

[9] Geoffrey E Hinton. 1990. Connectionist learning procedures. In *Machine learning*. Elsevier, 555–610.

[10] W Evan Johnson, Cheng Li, and Ariel Rabinovic. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 1 (2007), 118–127.

[11] Samantha Johnston. 2013. The prevalence of chronic fatigue syndrome/myalgic encephalomyelitis: a meta-analysis. *Clinical epidemiology* 5 (2013), 105.

[12] Winston Patrick Kuo et al. 2002. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18, 3 (2002), 405–412.

[13] Cosmin Lazar et al. 2013. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics* 14, 4 (2013), 469–490.

[14] Chinh Bkrong Nguyen et al. 2017. Whole blood gene expression in adolescent chronic fatigue syndrome: an exploratory cross-sectional study suggesting altered B cell differentiation and survival. *Journal of translational medicine* 15, 1 (2017), 102.

[15] Angela P Presson et al. 2008. Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC systems biology* 2, 1 (2008), 95.

[16] Michele Reyes et al. 2003. Prevalence and incidence of chronic fatigue syndrome in Wichita, Kansas. *Archives of internal medicine* 163, 13 (2003), 1530–1536.

[17] Matthew E Ritchie et al. 2015. LIMMA powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43, 7 (2015), e47–e47.

[18] Ana-Maria Šimundić. 2009. Measures of diagnostic accuracy: basic definitions. *Ejifcc* 19, 4 (2009), 203.

[19] Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. 2009. Multi-class adaboost. *Statistics and its Interface* 2, 3 (2009), 349–360.