2020-12

LightGWAS: A Novel Machine Learning Procedure for Genome-Wide Association Study

Ambrozio Bruno
Technological University Dublin

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Lucas Rizzo
Technological University Dublin

Recommended Citation

Bruno Ambrozio, Luca Longo, Lucas Rizzo. LightGWAS: A Novel Machine Learning Procedure for Genome-Wide Association Study, Proceedings for the 28th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, December 7-8, 2020, V. 2271, pp. 25-36, DOI: 10.6084/m9.figshare.13483341.v1

LightGWAS: A Novel Machine Learning Procedure for Genome-Wide Association Study

Conference Paper · December 2020

DOI: 10.6084/m9.figshare.13483341.v1

3 authors:

Bruno Ambrozio
Technological University Dublin - City Campus
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Luca Longo
Technological University Dublin - City Campus
90 PUBLICATIONS **1,086** CITATIONS
SEE PROFILE

Lucas Rizzo
Technological University Dublin - City Campus
22 PUBLICATIONS **144** CITATIONS
SEE PROFILE

LightGWAS: A Novel Machine Learning Procedure for Genome-Wide Association Study

Bruno Ambrozio[0000−0002−6180−6986], Luca Longo[0000−0002−2718−5426], and Lucas Rizzo[0000−0001−9805−5306]

School of Computer Science, Technological University Dublin, Ireland
d16128063@mytudublin.ie,{luca.longo,lucas.rizzo}@tudublin.ie

Abstract. This paper proposes a novel machine learning procedure for genome-wide association study (GWAS), named LightGWAS. It is based on the LightGBM framework, in addition to being a single, resilient, autonomous and scalable solution to address common limitations of GWAS implementations found in the literature. These include reliance on massive manual quality control steps and specific GWAS methods for each type of dataset morphology and size. Through this research, LightGWAS has been contrasted against PLINK2, one of the current state-of-the-art for GWAS implementations based on general linear model with support to firth regularisation. The mean differences measured upon standard classification metrics, extracted via quantitative empirical tests through k-fold cross-validation technique, indicated that LightGWAS outperforms PLINK2 for balanced, imbalanced, and high-imbalanced genomic datasets. Paired difference tests denoted statistical significance in the results extracted from the experiments with imbalanced datasets. This article contributes to the body of knowledge by presenting a potentially more efficient GWAS procedure based on nonparametric approaches. LightGWAS ensures adaptability with higher precision in the discovery of causal single-nucleotide polymorphisms, thanks to the leaf-wise tree growth algorithm offered by the state-of-the-art for gradient boosting decision trees. Control for false-positives and statistical power are automatically addressed by the model's training process, which significative reduces human dependency during the study design.

Keywords: LightGWAS, LightGBM, genome-wide association study.

1 Introduction

The most common type of genetic variant among humans' DNA is the single-nucleotide polymorphism (SNP) [22]. SNPs are responsible for phenotypes: observable characteristics or traits in a cohort [7]. Phenotypes can be modelled quantitatively, such as people's height, weight, body mass index, or blood pressure. Alternatively, they can be qualitative such as eye colour, curly hair, or a disease status like affected or not by Type-2 diabetes. Whenever a SNP is responsible for a phenotype, it is denominated as a causal-SNP. Therefore, identifying causal-SNPs is an effective way to understand, prevent, or treat complex illnesses.

There are many methods to discovery causal-SNPs, including genome-wide association study (GWAS). GWAS implementations calculate the association between each SNP and the underlying phenotype throughout a statistical model. Therefore, GWAS is roughly analogue to, or a type of feature selection: each SNP is a feature (independent variable), and the phenotype is the class (target, or dependent variable). The features identified as better predictors of the class are the potential causal-SNPs.

Statistical regression models portray the state-of-the-art for GWAS. Despite their efficiency, some eminent problems have become inevitable over the past years. These include reduction of costs for DNA sequencing [18], which in turn allowed an exponential growth of data; expansion of SNPs datasets that have contributed to overwhelming sparsity, with millions of SNPs, and few patients [13]; high-disperse (or high-dimensional) datasets, compromising the approaches available for GWAS as they are derived from linear (parametric) models [11]. Another point of concern emerges with imbalanced ratios of rare cases and several controls. Such a scenario tends to inflate false-positives when data is exploited by regression over qualitative features. [26]. Nowadays, these obstacles are addressed via several manual quality control steps to increase statistical power and avoid type 1 errors [10, 19, 21, 22, 24]. However, as much as data grows, so does the dependency on manual intervention. Hence, it opens margins for human mistakes and compromises the scalability of the study. To address the aforementioned gaps, this paper proposes a novel procedure for GWAS assembled over decision trees (DT) enhanced by gradient boosting machine (GBM), whose implementation comes from the LightGBM framework [9]. It ensures adaptability to the most diverse genomic data structures by controlling bias and variance over the training process. Consequently, it improves precision, independently of human intervention. Such a procedure has been named LightGWAS. Therefore, this work attempts to answer the following research question:

– *Can LightGWAS be an alternative method to the state-of-the-art for genome-wide association studies based upon general linear models, by increasing statistical power on causal-SNP detection, and reducing the number of manual quality control steps?*

The research goals of this paper are: (a) to evaluate whether LightGWAS is a suitable GWAS method for qualitative phenotypes, according to a set of common metrics for classification problems; and (b) to assess if LightGWAS outperforms the available state-of-the-art for GWAS in terms of statistical power and precision. Finally, the remainder of this paper is organised as follows: Section 2 reviews researches on the state-of-the-art for genome-wide association studys along with an overview of the LightGBM framework. Section 3 introduces the design and a set of hypotheses for answering the research question. Section 4, in turn, presents the results with a discussion. Lastly, section 5 concludes the study, highlighting its contributions and possible future work.

2 Literature review and related work

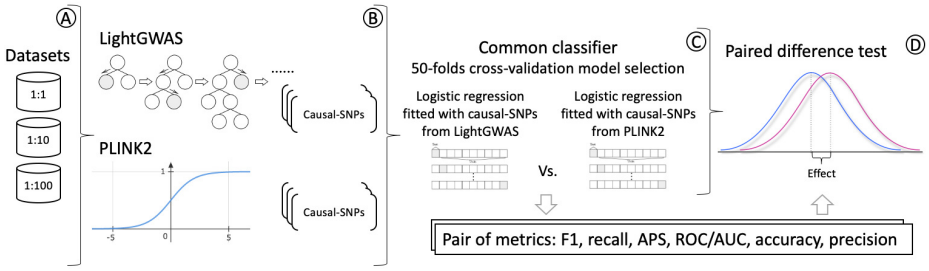
A Genome-wide association study (GWAS) is a discovery-driven research technique to catalogue single-nucleotide polymorphisms (SNPs) across populations and to identify genetic markers associated with traits [1, 4]. Since the completion of the human genome sequence in 2003, about 3,700 GWASs contributed to discovering thousands of genetic risk causal-SNPs and their biological functions [17, 15]. The state-of-the-art for GWAS methods are based on three exclusivelly statistical association models: general linear model (GLM), linear mixed model (LMM), and scalable and accurate implementation of generalized mixed model (SAIGE) [12]. Their applicability depends on the phenotype type, sample size, and cohort distribution across the manipulated genomic dataset. According to [12], the following criteria should be considered to select the appropriate model: (a) GLM implementations for quantitative traits, up to five thousand samples. If qualitative phenotype, the logistic regression implementation should include firth regularisation to minimize the fitting errors caused by the categorical class whenever its frequency is lower than 400 [14]; (b) LMM implementation for datasets bigger than five thousand samples and quantitative traits type. Whether qualitative phenotype, the dataset should be in a normal distribution; or (c) SAIGE [26] for high-imbalanced *case-control* ratio of qualitative traits. Independently of the chosen method, principal component analysis should also be employed. It helps to filter SNPs that might be caused by the structure of the population (generating confounding due to ancestry), rather than the investigated phenotype [19]. Usually, the first ten eigenvalues are arbitrarily considered as covariants for an association model [19, 2]. The GWAS outcome is a list of potential causal-SNPs.

This paper proposes a method for GWAS based on LightGBM [9]: a gradient boosted decision trees (GBDT) framework built upon histogram algorithms. LightGBM grows the trees leaf-wise and uses gradient-based one-side sampling (GOSS) to downsampling data, and exclusive feature bundling (EFB) to reduce feature dimension. In order to address GBDT problems related to high-computational complexity due to abundance of data, GOSS retains the large gradients samples, randomly selects small gradients, and assign constant weights to them. The algorithm concentrates on undertrained samples without altering the distribution of raw data. EFB, in turn, is a feature extraction technique, based on the graph coloring problem, which also contributes to reducing the histogram building complexity. It deals with the sparsity of the data by grouping many independent variables to the dense features, avoiding unnecessary computation with pieces that do not account for the outcome variable. LightGBM, within the proposal GWAS solution, discoveries causal-SNPs by calculating the model's feature importance. Each SNP is, in fact, an independent variable of the model. Hence, the list of features that better explains the dependent variable (phenotype) contains the saught SNPs. Considering the LightGBM framework design and the strong evidence of its inference to address problems involving high-sparse data over big datasets [16, 25, 23], this article works upon the idea that such a framework is also a potential core engine for GWAS.

3 Experiment design and methodology

LightGWAS is designed to be a GWAS procedure based on a machine learning nonparametric method. The solution is composed of a GBDT algorithm implemented by the LightGBM framework [9]. It is fitted with the SNPs as independent variables and the phenotype as the class. Thus, the causal-SNPs are retrieved by calculating the models' feature importance. In turn, to answer the research question of this paper, an experiment involving three datasets, two feature selector models and a predicting model is conducted. Fig. 1 depicts the experiment design in four steps, followed by the evaluation strategy applied.

Fig. 1: Diagrammatic visualisation of experiment design, components and evaluation.



The datasets (Fig. 1A) contain the same number of SNPs each, but varying the phenotype balance ratio on *cases:controls* of *1:1*, *1:10*, and *1:100*. The first two models (Fig. 1B) are GWAS procedures. One of the GWAS methods is the novelty behind this paper, the LightGWAS. The other one is PLINK2 [7], one of the state-of-the-art implementations for GWAS in contexts where GLM is required. Therefore, six causal-SNPs result sets are generated from them. The third model (Fig. 1C) referred to as *common classifier* from now on, is a logistic regression. It employes k-fold cross-validation model selector technique, with k been set arbitrarily to 50. A value higher than 30 was necessary to perform statistically significant comparisons across the resulting sets. The *common classifier* is fitted once with causal-SNPs discovered by LightGWAS as independent variables and another time with causal-SNPs retrieved with PLINK2. The dependent variable, in both circumstances, is the underlying phenotype of the datasets. Therefore, its output is a paired set of classification metrics from each of the GWAS methods. Lastly (Fig. 1D), the group of metrics extracted with the cross-validation are evaluated in terms of statistical significance for possible differences among them. The evaluated alternative hypotheses are:

– H_1**:** LightGWAS outperforms GLM based on logistic regression with firth regularisation for GWAS, across genomic datasets of balanced qualitative

phenotypes ($case : control = 1 : 1$), in terms of accuracy, precision, F1 score, and ROC/AUC.
- H_2**:** LightGWAS outperforms GLM based on logistic regression with firth regularisation for GWAS, across genomic datasets of imbalanced qualitative phenotypes ($case : control = 1 : 10$), in terms of precision, F1 score, and ROC/AUC.
- H_3**:** LightGWAS outperforms GLM based on logistic regression with firth regularisation for GWAS, across genomic datasets of high-imbalanced qualitative phenotypes ($case : control = 1 : 100$), in terms of precision, F1 score, and ROC/AUC.

3.1 Datasets

A GWAS relies on two different data groups: the genomic data that contains the DNA variances, and the traits to be associated with the SNPs between the *cases* and *controls* cohorts. Usually, the traits to be investigated are human phenotypes, such as diseases status, that can be retrieved from the patients electronic health records (EHR) [26]. In this article, selected datasets are fully synthetic in either genomic and phenotype data. Simulations have been introduced to distinguish accurately the causal-SNPs expected to be exposed by each of the evaluated GWAS models, which is paramount to compare them correctly. Dataset simulation for GWAS methods validation is a prevalent practice and can be observed in many types of researches, such as [5, 7, 8, 14, 26]. Accordingly, six datasets have been created, combined into three data groups of class (phenotype status) distribution: balanced, imbalanced, and high-imbalanced data. They have been named as *ds1_1*, *ds1_10* and *ds1_100*, respectivelly. The number of samples (fictitious patients) in each of the datasets respected the following pattern: *ds1_1 = case:control=1:1=2500:2500, N=5000*; *ds1_10 = case:control=1:10=400:4000, N=4400*; *ds1_100 = case:control=1:100=50:5000, N=5050*. The datasets were produced using the PLINK SNP simulation tool[1]. Each sample had a phenotype status class (*case* or *control*) and 10100 numeric features (each feature is a SNP). Further details about the variables of interest along with the parameters set to simulate the datasets can be found in Appendix A (table 2, page 12).

3.2 Procedure

The complete procedure to accomplish the objectives, and test the alternative hypotheses includes seven steps:

1. Simulation of datasets, as outlined above, in section 3.1.
2. LightGWAS implementation. It is composed of a GBDT implementation called LightGBM. The hyperparameters are tunned through 200 iterations of randomised 5-folds cross-validation search. Table 3 in the Appendix B (page 12) contains the cross-validated optimal hyperparameters selected for each dataset group.

[1] `http://zzz.bwh.harvard.edu/plink/simulate.shtml`

3. Discover the causal-SNPs across the early mentioned datasets by employing LightGWAS and PLINK2. Therefore, two sets of causal-SNPs per GWAS method is generated. PLINK's outcome is a set of SNPs accompanied by their p-value. The causal-SNPs filtering is reached by assuming a cut-off (α) for such a p-value. For the datasets *ds1_1* and *ds1_10*, the cut-off $p \leq \alpha | \alpha = 5 \times 10^{-8}$ is assumed, as per genome-wide association study convension [3]. In turn, for the dataset *ds1_100*, the cut-off is $p \leq \alpha | \alpha = 5 \times 10^{-4}$ because no SNP was selected with the first one. This decision has been grounded on [14]. In contrast, LightGWAS selects each SNP with the *gain* or *split* score of the decision trees. Therefore, the list of features importance from the LightGBM framework is the set of causal-SNPs retrieved with LightGWAS.
4. GWAS model's evaluation. In order to compare how effective LightGWAS is in comparison to PLINK, the *common classifier* is employed. It is a logistic regression executed through 50-folds cross-validation for model selection, which is fitted upon two conditions: one with the features as the causal-SNPs collected via LightGWAS, and another with causal-SNPs selected via PLINK. The class (or target) for both scenarios, is the phenotype variable. Therefore, the *common classifier* output is a separated dataset with 50 result samples per GWAS model. The following metrics have been evaluated: weighted average of the precision and recall (F1), recall, average precision score (APS), receiver operating characteristic (ROC)/area under the curve (AUC), accuracy, and precision.
5. The confidence interval (CI) of the metric's result sets are calculated through 5000 bootstraps in a cut-off of $\alpha = 0.05$. The subsamples (resampling with replacement) is sized at 50% ($N \times 0.5$). Therefore, there is 95% of a likelihood that the reported lower limit (LL) and upper limit (UL) represent the confidence intervals of the true metrics' performances.
6. Paired difference tests are employed to measure how significant is the observed differences in each metric pair. Dependent (paired) sample Student's t-test is applied to the metric pairs that held a normal distribution, and Wilcoxon signed-rank test otherwise. Tests to assess whether a metric (variable of the results dataset) is in a Gaussian distribution are conducted with D'Agostino's K^2 Normality Test. Whenever a sample does not reach the levels of a normal distribution, power transformation through Box-cox is firstly attempted before assuming nonparametric approaches.
7. The effect of the observed mean differences are calculated through Cohen's d test when the parametric test has been used, and Wilcoxon r score otherwise.

4 Results and evaluation

The consolidated results can be observed below in table 1, followed by the statistical report. The CI ranges along with the standard deviation (SD) of each metric have been logged to the Appendix C (table 4, page 12).

Table 1: Results of statistical tests. (§) Metric's "p-value" and "stat" calculated from the Box-Cox power transform result. (*) Metric statistically significant on $\alpha = 0.05$. (**) Metric statistically significant on $\alpha = 0.01$. (MD) mean absolute difference. Best values in bold.

		LightGWAS (Mean)	**PLINK (Mean)**	**MD**	**Stat**	**p-value**	**Effect**
ds1_1	f1	**0.967436**	0.967 416	0.000 020	$2.879\,656 \times 10^{-2}$	0.977 144	0.001 191
	recall	**0.966800**	0.966 400	0.000 400	$3.747\,014 \times 10^{-1}$	0.709 499	0.019 789
	APS§	0.995 725	**0.995748**	0.000 022	0.559 679	0.578 248	0.006 192
	ROC/AUC§	**0.995664**	0.995 648	0.000 016	0.744 993	0.459 835	0.004 531
	accuracy	0.967 400	0.967 400	0	$3.172\,727 \times 10^{-15}$	1	0
	precision	0.968 505	**0.968896**	0.000 390	$-4.497\,929 \times 10^{-1}$	0.654 843	0.015 893
ds1_10	f1*	**0.993251**	0.991 394	0.001 857	2.364 684	**0.022051**	0.292 229
	recall	**0.993750**	0.993 000	0.000 750	113.5	0.662 096	**16.051324**
	APS**	**0.999830**	0.999 671	0.000 159	54.0	**0.002024**	7.636 753
	ROC/AUC**	**0.998281**	0.996 719	0.001 562	48.5	**0.006190**	6.858 936
	accuracy*	**0.987727**	0.984 318	0.003 409	2.393 172	**0.020579**	0.294 840
	precision**	**0.992842**	0.989 887	0.002 955	37.5	**0.006574**	5.303 301
ds1_100	f1	**0.997205**	0.996 713	0.000 492	183.0	0.430 596	**25.880108**
	recall	**0.998600**	0.999 400	0.000 800	5.0	0.234 194	**0.707107**
	APS	**0.999857**	0.999 823	0.000 034	163.5	0.095 638	**23.122392**
	ROC/AUC	**0.987000**	0.982 600	0.004 400	166.5	0.107 381	**23.546656**
	accuracy	**0.994455**	0.993 465	0.000 990	180.0	0.387 660	**25.455844**
	precision	**0.995830**	0.994 053	0.001 776	176.0	0.342 925	**24.890159**

4.1 Statistical report

Below follows a statistical report, separated by dataset group, extracted from the interpretation of the consolidate result sets disclosed in table 1.

Dataset ds1_1: LightGWAS slightly outperformed PLINK on metrics *F1*, *recall*, and *ROC/AUC*, while PLINK outperformed LightGWAS on *APS*, and *precision*. Both models reached out the same mean value for *accuracy* so that zero mean absolute difference (MD). The t-tests indicated no statistical significance on $\alpha = 0.05$ for any of the measured metrics. The standardized difference between the means resulted in a small effect for all of the metrics ($d < 0.5$). In terms of causal-SNP selection, LightGWAS selected 86 SNPs, while PLINK selected 90. PLINK managed to pick all SNPs selected by LightGWAS, plus other four causal-SNPs.

Dataset ds1_10: LightGWAS slightly outperformed PLINK for every measured metrics. The t-tests indicated statistical significance on $\alpha = 0.05$ for both *F1* and *accuracy* with small effect ($d < 0.5$). The Wilcoxon test indicated statistical significance on $\alpha = 0.01$ and large effect ($r \geq 0.8$) for *APS*, *ROC/AUC* and *precision*. No statistical significance on $\alpha = 0.05$ has been observed for *recall*, although the observed difference had large effect ($r \geq 0.8$). In terms of causal-SNP selection, LightGWAS selected 80 SNPs, while PLINK selected 76. LightGWAS managed to pick all SNPs selected by PLINK, plus other four causal-SNPs.

Dataset ds1_100: LightGWAS slightly outperformed PLINK for every measured metrics. The Wilcoxon test indicated no statistical significance on $\alpha = 0.05$ for any of them. However, a medium effect ($r \geq 0.5 \wedge r < 0.8$) has been observed for *recall*, and a large effect ($r \geq 0.8$) for all the other metrics. In terms of causal-SNP selection, LightGWAS selected 28 SNPs, while PLINK selected 19. LightGWAS managed to pick 14 SNPs missed by PLINK, and PLINK, in turn, managed to select 5 SNPs missed by LightGWAS.

4.2 Discussion

The models implemented through LightGWAS performed as good as PLINK for GWAS over the balanced dataset. The paired difference tests disclosed that none of the measured differences is statistically significant on cut-off $\alpha = 0.05$. Also, the observed effects through Cohen's d presented a small standardised effect between all the means of the paired metrics. Consequently, the alternative hypothesis H_1 had to be rejected as LightGWAS did not outperform (neither underperformed) statistically significant for such a dataset.

The experiments involving an imbalanced dataset brought evidence that supports accepting the alternative hypothesis H_2. LightGWAS has outperformed PLINK for such a scenario. Although *recall* did not reach statistical significance on $\alpha = 0.05$ (therefore as good as PLINK), all the other metrics had relevant results on $\alpha = 0.01$ (*F1* and *accuracy* on $\alpha = 0.05$). Furthermore, the metrics measured through nonparametric tests (*recall*, *APS*, *ROC/AUC* and *precision*) resulted in a large effect ($r \geq 0.8$).

The alternative hypothesis H_3 was rejected. Although LightGWAS outperformed PLINK with medium effect for *recall* ($r \geq 0.5 \wedge r < 0.8$) and a large effect for the other metrics ($r \geq 0.8$) when instantiated with a high-imbalanced dataset, none of the results reached statistical significance on $\alpha = 0.05$.

Considering exclusively the k-fold cross-validation model selection results (observed differences in the means), models implemented via the proposed LightGWAS procedure outperformed those implemented with PLINK in the three evaluated scenarios. However, if taking into consideration the statistical analysis of the metrics pairs differences, this result is held with statistical significance only in the experiments involving the imbalanced dataset. Nonetheless, according to [6], it is important to note that statistical significance should not be the exclusive approach to reject how relevant a model is. The scientific perspective (or significance) of the underlying problem should also be taken into consideration. Genome-wide association study plays an essential rule on identifying causal anomalies across DNA, and any improvement over a method, being it statistically significant or not, should be accounted for. Hence, although some of the measured metrics did not reach statistical significance (leading to the rejection of the alternative hypotheses H_1 and H_3), they prove to be scientifically meaningful through their effect differences and the number of discovered causal-SNPs. As a result, the research question (page 1) can be answered positively. The evidence collected from the tested hypotheses supports the theory that LightGWAS is a potential genome-wide association study method.

5 Conclusion

This paper has proposed a novel genome-wide association study (GWAS) procedure, named LightGWAS. It is a nonparametric machine learning (ML) method based on the LightGBM framework [9]. LightGWAS has been idealised as a potential single, resilient, autonomous and scalable solution to address some of the found limitations of the available state-of-the-art implementations for GWAS. A literature review identified that the current GWAS implementations rely on cumbersome manual quality control steps to address statistical problems, such as controlling for false-positive inflation and power reduction. These challenges increase as the data grows or becomes imbalanced. It also showed they demand a particular GWAS method for each type of genomic data structure, which increases human dependency. In this research, the effectiveness of the models implemented via the proposed LightGWAS procedure was assessed upon GWAS scenarios where the investigated phenotype is qualitative and datasets are about to five thousand samples of balanced ($case : control = 1 : 1$), imbalanced ($case : control = 1 : 10$), and high-imbalanced ($case : control = 1 : 100$) genomic data. Next, LightGWAS models were contrasted with those implemented via the state-of-the-art for GWAS (PLINK2 [7]). This assessment was performed through an empirical comparative experiment. A model selection based on 50-fold cross-validation signed out LightGWAS as the best choice in terms of mean differences. The results from empirical statistical tests denoted that the differences are statistically significant for imbalanced datasets contexts.

The main contribution of LightGWAS for genome-wide association study is the fact it is based on a nonparametric machine learning approach against the state-of-the-art that strongly relies on parametric statistical models. Therefore, LightGWAS allows scalability and adaptability to the most diverse genomic data morphology, which, in turn, reduces human dependency. It scales thanks to the LightGBM framework, which is the state-of-the-art for gradient boosted decision trees, capable of handling large and high-sparse datasets. LightGBM was created to address classification or regression problems. Still, in the LightGWAS procedure, it is used as a phenotype causal single-nucleotide polymorphism (SNP) discover by calculating the feature importance of a fitted model. Hence, this research shows originality by taking a specific technique and adapting it to a new domain of application. For all these reasons, LightGWAS is a new contribution from data science towards the evolvement of molecular biology science.

For future work, it is recommended to compare LightGWAS with the GWAS procedures based on linear mixed model, and scalable and accurate implementation of generalized mixed model. Thus, the effectiveness of LightGWAS can also be assessed against scenarios that go beyond the ones addressable through general linear models. It would also benefit whether using quantitative phenotypes to make sure LightGWAS attends to linear association models. Lastly, it is recommended the development of a mechanism to identify causal-SNPs from decision trees *gain or split* scores, as no p-values exist in such a context. It is crucial to develop a system analogue to the cut-offs employed by the current state-of-the-art regression models to filter causal-SNPs ($p \leq \alpha$ for each SNP).

References

1. Bush, W.S., Moore, J.H.: Chapter 11: Genome-wide association studies. PLoS Computational Biology **8**(12), e1002822 (Dec 2012). https://doi.org/10.1371/journal.pcbi.1002822, https://doi.org/10.1371/journal.pcbi.1002822
2. Chen, X., Ishwaran, H.: Random forests for genomic data analysis. Genomics **99**(6), 323–329 (Jun 2012). https://doi.org/10.1016/j.ygeno.2012.04.003, https://doi.org/10.1016/j.ygeno.2012.04.003
3. Fadista, J., et al.: The (in)famous GWAS p-value threshold revisited and updated for low-frequency variants. European Journal of Human Genetics **24**(8), 1202–1205 (Jan 2016). https://doi.org/10.1038/ejhg.2015.269, https://doi.org/10.1038/ejhg.2015.269
4. Farrell, R.E.: Functional genomics and transcript profiling. In: RNA Methodologies, pp. 685–695. Elsevier (2017). https://doi.org/10.1016/b978-0-12-804678-4.00024-5, https://doi.org/10.1016/b978-0-12-804678-4.00024-5
5. Golan, D., Rosset, S., Lin, D.Y.: Mixed models for case-control genome-wide association studies: Major challenges and partial solutions. In: Handbook of Statistical Methods for Case-Control Studies, pp. 495–514. Chapman and Hall/CRC (Jun 2018). https://doi.org/10.1201/9781315154084-27, https://doi.org/10.1201/9781315154084-27
6. Greenland, S., et al.: Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology **31**(4), 337–350 (Apr 2016). https://doi.org/10.1007/s10654-016-0149-3, https://doi.org/10.1007/s10654-016-0149-3
7. Hill, A., Loh, P.R., Bharadwaj, R.B., Pons, P., Shang, J., Guinan, E., Lakhani, K., Kilty, I., Jelinsky, S.A.: Stepwise distributed open innovation contests for software development: Acceleration of genome-wide association analysis. GigaScience **6**(5) (Feb 2017). https://doi.org/10.1093/gigascience/gix009, https://doi.org/10.1093/gigascience/gix009
8. Jiang, L., et al.: A resource-efficient tool for mixed model association analysis of large-scale data. Nature Genetics **51**(12), 1749–1755 (Nov 2019). https://doi.org/10.1038/s41588-019-0530-8, https://doi.org/10.1038/s41588-019-0530-8
9. Ke, G., et al.: Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 3146–3154. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf
10. Lee, S., Wright, F.A., Zou, F.: Control of population stratification by correlation-selected principal components. Biometrics **67**(3), 967–974 (Dec 2010). https://doi.org/10.1111/j.1541-0420.2010.01520.x, https://doi.org/10.1111/j.1541-0420.2010.01520.x
11. Li, J., et al.: Feature selection. ACM Computing Surveys **50**(6), 1–45 (Jan 2018). https://doi.org/10.1145/3136625, https://doi.org/10.1145/3136625
12. Loh, P.R., et al.: Mixed-model association for biobank-scale datasets. Nature Genetics **50**(7), 906–908 (Jun 2018). https://doi.org/10.1038/s41588-018-0144-6, https://doi.org/10.1038/s41588-018-0144-6
13. Lubke, G., et al.: Gradient boosting as a SNP filter: an evaluation using simulated and hair morphology data. Journal of Data Mining in Genomics & Proteomics

04(04) (2013). https://doi.org/10.4172/2153-0602.1000143, https://doi.org/10.4172/2153-0602.1000143
14. Ma, C., et al.: Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genetic Epidemiology **37**(6), 539–550 (Jun 2013). https://doi.org/10.1002/gepi.21742, https://doi.org/10.1002/gepi.21742
15. Mills, M.C., Rahal, C.: A scientometric review of genome-wide association studies. Communications Biology **2**(1), 9 (Jan 2019). https://doi.org/10.1038/s42003-018-0261-x, https://doi.org/10.1038/s42003-018-0261-x
16. Mo, K., Li, J.: A deep auto-encoder based LightGBM approach for network intrusion detection system. In: Proceedings of the International Conference on Advances in Computer Technology, Information Science and Communications. pp. 142–147. SCITEPRESS - Science and Technology Publications (2019). https://doi.org/10.5220/0008098401420147, https://doi.org/10.5220/0008098401420147
17. Pearson, T.A.: How to interpret a genome-wide association study. JAMA **299**(11), 1335 (Mar 2008). https://doi.org/10.1001/jama.299.11.1335, https://doi.org/10.1001/jama.299.11.1335
18. Pérez-Enciso, Zingaretti: A guide for using deep learning for complex trait genomic prediction. Genes **10**(7), 553 (Jul 2019). https://doi.org/10.3390/genes10070553, https://doi.org/10.3390/genes10070553
19. Price, A.L., et al.: Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38**(8), 904–909 (Jul 2006). https://doi.org/10.1038/ng1847, https://doi.org/10.1038/ng1847
20. Purcell, S., et al.: PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics **81**(3), 559–575 (Sep 2007). https://doi.org/10.1086/519795, https://doi.org/10.1086/519795
21. Reed, E., et al.: A guide to genome-wide association analysis and post-analytic interrogation. Statistics in Medicine **34**(28), 3769–3792 (Sep 2015). https://doi.org/10.1002/sim.6605, https://doi.org/10.1002/sim.6605
22. Sebastiani, P., et al.: Genome-wide association studies and the genetic dissection of complex traits. American Journal of Hematology **84**(8), 504–515 (Aug 2009). https://doi.org/10.1002/ajh.21440, https://doi.org/10.1002/ajh.21440
23. Song, Y., et al.: Prediction of double-high biochemical indicators based on LightGBM and XGBoost. In: Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science - AICS 2019. p. 189–193. ACM Press (2019). https://doi.org/10.1145/3349341.3349400, https://doi.org/10.1145/3349341.3349400
24. Spencer, C.C.A., et al.: Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genetics **5**(5), 1–13 (May 2009). https://doi.org/10.1371/journal.pgen.1000477, https://doi.org/10.1371/journal.pgen.1000477
25. Wang, R., et al.: Power system transient stability assessment based on bayesian optimized LightGBM. In: 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2). pp. 263–268. IEEE (Nov 2019). https://doi.org/10.1109/ei247390.2019.9062027, https://doi.org/10.1109/ei247390.2019.9062027
26. Zhou, W., other: Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics **50**(9), 1335–1341 (Aug 2018). https://doi.org/10.1038/s41588-018-0184-y, https://doi.org/10.1038/s41588-018-0184-y

Appendices

Appendix A: Datasets' phenotype ratios and variables of interest

Table 2: Phenotype ratios for genetic datasets build-up (top), and variables of interest extracted from the executed simulations (bottom). Values have been based on the PLINK SNP simulation tool documentation [20]. Due to space limitations, the concepts of minor allele frequency (MAF), heterozygotes, and homozygotes are not expanded. However, they can be consulted at [1, 15, 20].

no. SNPs	SNP Prefix	Lower allele frequency	Upper allele frequency range	Odds ratio for heterozygotes	Odds ratio for homozyygotes
10000	n	0.00	1.00	1.00	1.00
100	d	0.00	1.00	2.00	4.00

Variable	Type	Range	Sample
Individual ID	Nominal	Alphanumeric	*per*13
Phenotype	Numeric	1=control, 2=case	2
...			
n_1351_T(/A) d_13_G(/T)	Numeric	[0, 1 or 2]	2
...			

Appendix B: LightGBM hyperparameter values

Table 3: LightGBM parameters selected via 200 iterations of randomised 5-folds cross-validation.

	ds1_1	ds1_10	ds1_100
colsample_bytree	0.47328041	0.47328041	0.866621446
learning_rate	0.03	0.03	0.01
max_depth	1	1	6
min_child_samples	147	147	454
min_child_weight	1.0	1.0	1.0
min_split_gain	0	0	0
n_estimators	2000	2000	2000
num_leaves	35	35	41
reg_alpha	0.1	0.1	5
reg_lambda	0.1	0.1	50
subsample	0.995930118	0.995930118	0.820421212
subsample_for_bin	200000	200000	200000

Appendix C: Confidence interval ranges and standard deviations

Table 4: Bootstrap 95% confidence interval (CI) metric ranges and standard deviations (SDs).

		LightGWAS			PLINK		
		SD	LL	UL	SD	LL	UL
ds1_1	f1	0.017 298	0.961 616	0.981 966	0.016 862	0.961 767	0.983 936
	recall	0.020 045	0.952 000	0.984 000	0.020 380	0.952 000	0.984 000
	APS	0.003 669	0.994 011	0.998 711	0.003 506	0.994 256	0.998 870
	ROC/AUC	0.003 572	0.994 760	0.998 672	0.003 490	0.993 080	0.998 848
	accuracy	0.017 474	0.962 000	0.982 000	0.017 001	0.962 000	0.984 000
	precision	0.024 702	0.963 563	0.987 904	0.024 434	0.963 710	0.991 701
ds1_10	f1	0.005 909	0.987 562	0.996 255	0.006 772	0.985 000	0.993 789
	recall	0.009 193	0.990 000	1.000 000	0.009 161	0.985 000	0.997 500
	APS	0.000 272	0.999 462	0.999 925	0.000 486	0.999 183	0.999 863
	ROC/AUC	0.002 738	0.994 750	0.999 250	0.004 748	0.991 938	0.998 625
	accuracy	0.010 729	0.977 273	0.993 182	0.012 340	0.972 727	0.988 636
	precision	0.008 637	0.980 344	0.995 000	0.010 093	0.980 247	0.992 537
ds1_100	f1	0.003 806	0.994 024	0.997 509	0.002 956	0.994 000	0.997 009
	recall	0.004 522	0.996 000	1.000 000	0.002 399	0.994 000	1.000 000
	APS	0.000 565	0.999 624	0.999 984	0.000 304	0.999 330	0.999 851
	ROC/AUC	0.048 498	0.964 000	0.998 400	0.029 264	0.937 600	0.985 200
	accuracy	0.007 527	0.988 119	0.994 759	0.005 869	0.988 119	0.994 059
	precision	0.004 951	0.990 079	0.996 008	0.004 905	0.990 079	0.995 036