Studying the Acquisition Function of Bayesian Optimization With Machine Learning With DNA Reads

Jacob Porter
Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061

***Abstract*—Bayesian optimization is a method of minimizing a function can be evaluated but the function cannot be explicitly written. This is useful for hyperparameter tuning in machine learning. Minimizing the loss on the validation data is a function where the hyperparameters of the machine learning model are the input to the function. I study the choice of the acquisition function in optimization performance using machine learning with random forests and multilayer perceptrons. The acquisition function is used to choose the next location for function evaluation as function evaluation is done sequentially. The machine learning task involves predicting read mapping performance of bisulfite-treated short DNA reads based on features extracted from the DNA reads. Bisulfite-treatment is a sequencing method that is used to study the epigenetic methylation of cytosine nucleic acids. The results of Bayesian optimization were similar to grid search, and there was little difference found in the choice of acquisition function. The implementation of Bayesian optimization from skopt used parallelization less efficiently than grid search.**

1. Introduction

A DNA read sequencer produces short DNA fragments from an organism, and DNA sequence alignment maps these short DNA reads, which are strings over the nucleic acid bases A, C, T, and G, to a reference genome. This process can be error prone as the short DNA fragments may not match a portion of the reference genome perfectly because of natural variation and mutation or because of sequencing error. Insight into why DNA mapping and alignment fails could lead to more effective alignment software. I use machine learning with features taken from the short DNA fragments to predict which reads will align well.

A challenging read mapping task involves epigenetic cytosine covalent modification. Epigenetic phenomena are heritable biology that does not come from DNA sequence data [1]. One of the most important and well studied epigenetic phenomena is the covalent modification of the cytosine nucleic acid. The 5-carbon of cytosine can be covalently bonded to a hydroxymethyl, methyl, formyl, or carboxylic group. The epigenetic methylation of cytosine plays an important role in disease, development, and gene regulation. Life experiences such as stress and toxin consumption affect epigenetic phenomena in heritable ways.

One way to identify the locations of DNA methylation is to sequence the DNA of an organism after it has been treated with bisulfite and then to identify nucleic acid base locations on a reference genome that differ in such a way as to suggest covalent modification of the cytosine base. Bisulfite converts unmethylated cytosine into thymine after polymerase chain reaction (PCR) amplification. Bisulfite treatment introduces more variation between the short DNA reads and the reference genome, so alignment tasks with bisulfite-treated DNA can be characterized by low alignment quality [2].

DNA sequence mapping software that is used for regular untreated reads includes Bowtie2 [3], BWA [4], and BFAST [5]. Mapping software for bisulfite-treated reads must adjust for the bisulfite treatment, and such software includes Bismark [6], BWA-Meth [7], and BisPin [8]. There are many more examples of these kinds of software. I used BisPin to map bisulfite-treated reads since it performed well in a previous study, and then I used machine learning with scikit learn to find features of the DNA reads that correspond with good alignment performance.

This project was used to explore Bayesian optimization for machine learning hyperparameter tuning. Bayesian optimization is a way of finding a minimum of an objective function when the function cannot be written explicitly but it can be evaluated [9]. Bayesian optimization itself has tuneable hyperparameters, and these include the acquisition function and the covariance function. The acquisition function is used to choose which function point to evaluate next, and the covariance kernel function determines how the function evaluation observations are correlated and the probabilities of unobserved function evaluations are calculated. I study the choice of acquisition function.

2. Related Work

Bayesian optimization for machine learning is discussed in the paper [9]. The paper claims that the expected improvement acquisition function and a Matern kernel work the best for machine learning hyperparamter tuning, but no evidence or research method for this is given. My study attempts to address this by empirically evaluating Bayesian

optimization for hyperparameter tuning on a data set by varying the choice of acquisition function.

Other work has used machine learning to predict methylation loci from DNA reads [10], [11], [12], DNA age from methylation [13], [14], and DNA function from DNA sequence identity [15], but I could not find examples of research for predicting read alignment quality with machine learning. My own study found that entropy corresponds to read alignment categories [16].

3. Methods

3.1. Data Generation

Two million bisulfite-treated 200 base pair reads from SRA number SRR1104850 was downloaded from the Sequence Read Archive [17]. These reads were aligned to the human reference genome with BisPin on its default settings without any ambiguously mapped read rescoring. Reads with ambiguous bases (N's) were removed from the analysis since they would create a mixed single with entropy, one of the features.

Sixty features were created from the DNA short reads. These features included Shannon entropy [18], monomer frequency, average read quality, average read quality difference, read quality variance, skewness, and kurtosis, two sequence complexity measures from the paper [19], and features for run lengths. The read was split into thirds, and similar features for each third were created. Quality kurtosis features and the read length feature were removed since they were found to be uninformative. This left 55 features for prediction. Entropy has been used in the DNA read quality trimmer InfoTrim [20] and the bisulfite read mapper BatMeth [21].

Each read was assigned a read alignment category from the BisPin read mapper. These categories were uniquely mapped, ambiguously mapped, filtered, and unmapped. Previous work found that entropy corresponded to alignment categories [16]. This data is characterized by rare or imbalanced classes. Since of one million reads, only a few thousand are in the ambiguously mapped or unmapped categories.

3.2. Machine Learning

Sci-kit learn (v0.19.1) was used to perform machine learning, and scikit-optimize implemented Bayesian optimization (https://scikit-optimize.github.io/). One million reads were used to do 3-fold cross validation, and then 500k reads were used as a test set to determine the model's performance with the accuracy metric. The features were standardized with a `StandardScaler`, and `GridSearchCV` was used for a grid search of the hyperparameters. `BayesSearchCV`, which uses Bayesian optimization for hyperparameter tuning from scikit-optimize, was used as a drop in replacement for `GridSearchCV`. `BayesSearchCV` uses a Matern kernel to determine covariance, and the parameters for the kernel are autotuned as Bayesian optimization progresses. There were four acquisition functions tested: expected improvement (EI), lower confidence bound (LCB), negative probability of improvement (PI), and a hedging function that computes all of the available acquisition functions and uses the softmax of them (hedge). There are two other acquisition functions that incorporate training time, but these were not tested. In Bayesian optimization, the number of iterations are the number of function evaluations performed.

Acquisition Function	Best Parameters	Test set accuracy
EI	16,1.0	77.9826
LCB	15,1.0	77.9782
PI	14,1.0	77.9428
Hedge	16,1.0	77.9126

Figure 1. Random forest models with 25 iterations and the optimial max depth and max features with test set accuracy.

Two machine learning models were used. Random forests (RF) were trained, and the hyperparameters max depth and max features were optimized. There are several other hyperparameters for random forests. Multi-layer perceptrons (MLP) with the regularization hyperparameter alpha was optimized. The ReLU activation function with stochastic gradient descent and an adaptive learning rate was used for the MLPs. An initial optimization of the number and depth of the layers was done, and four hidden layers with depths (30, 20, 15, 10) were found to be optimal.

The machine learning analysis was done on an Intel Core i7-5820K six core CPU @ 3.30 GHz with 48 GB RAM, and timing information was calculated with Python's `now` function from the `datetime` library.

4. Results

The random forest models were trained with the four acquisition functions with 25 function evaluations, and the results are shown in Figure 1. Each model learned approximately the same best parameters with some small differences in the max depth. The differences in accuracy are a result of the random bootstraps that random forests produce rather than the choice of acquisition function since the acquisition functions are choosing almost identical hyperparameter settings. The total training time for each acquisition function was approximately 1 hour and 20 minutes.

Grid search and Bayesian optimization was compared with 110 iterations each in Figure 2. Grid search took approximately 4 hours to complete with 5 processes, and Bayesian optimization took 4 hours and 50 minutes to complete but it was only able to use 3 processes at a time because of the 3-fold cross validation and the fact that Bayesian optimization is implemented to sample one function evaluation at a time. The hedge acquisition function was used, and the test set accuracy was similar to grid search. The test set accuracy of these executions is slightly better than those given in Figure 1, and this is probably because of 110 models, some of the bootstrap replicates will happen to be more predictive.

Acquisition Function	Best Parameters	Test set accuracy
Grid Search	15,0.6	78.1394
Hedge	15,0.4807802	78.0888

Figure 2. Random forest models with 110 iterations for grid search and Bayesian optimization. Optimal max depth and max features with test set accuracy are shown.

Acquisition Function	Best Parameters	Test set accuracy
Grid Search	1.0	77.2996
Hedge	0.252171	77.278

Figure 3. MLP, 10 iterations, best alpha.

The results in Figures 1 and 2 suggest that using Bayesian optimization with 25 executions is competitive in accuracy and better in training time than the more exhaustive grid search of 110 executions.

Figure 3 shows results for the MLP models with only 10 executions. The tuned alpha hyperparameter is quite different, but the test set accuracy is similar. Both executions were performed with 3 processes to determine the computational overhead of using Bayesian optimization. The Bayesian optimization execution took 7 minutes 37 seconds, and the grid search execution took 7 minutes 20 seconds. This suggests that computational overhead for Bayesian optimization is small averaging 1-2 seconds per execution.

5. Conclusions

Hyperparameter tuning with Bayesian optimization is competitive with grid search, and little evidence was found to prefer one acquisition function over another.

Perhaps Bayesian optimization with fewer executions and much less training time will be as optimal as high dimensional grid searches, but in low dimensional grid searches, grid search may be more optimal in time because it exploits parallelization more efficiently in the `GridSearchCV` and `BayesSearchCV` implementations. A better use of parallelization for Bayesian optimization may be possible using Monte Carlo simulation as discussed in the paper [9], and this could make Bayesian optimization as competitive in time with multiple processes as grid search.

Perhaps floating point hyperparameters are more effectively optimized than discrete hyperparameters with Bayesian optimization, and future work could include exploring this question. Future work could also include exploring the effects of different covariance kernels on optimization.

Maybe formal mathematical criterion could be developed to give reasons why some kernel or acquisition function would be better for machine learning hyperparameter tuning. Are close hyperparameters produce generally smooth, continuous functions for some reason?

This prediction problem can be researched in its own right by performing feature selection and by dealing with the imbalanced classes problem. Features predictive of mapper performance could lead to insights into producing a better read mapper, and the imbalanced class problem could be addressed by framing the problem as a regression problem where the alignment score is used as a response variable instead of alignment category. A variety of data can be examined including regular DNA reads.

References

[1] C. D. Allis, T. Jenuwein, D. Reinberg, and M.-L. Caparros, *Epigenetics*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY, 2007.

[2] H. Tran, J. Porter, M.-a. Sun, H. Xie, and L. Zhang, "Objective and comprehensive evaluation of bisulfite short read mapping tools," *Advances in Bioinformatics*, vol. 2014, p. 11, 2014.

[3] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, p. 357, 2012.

[4] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[5] N. Homer, B. Merriman, and S. F. Nelson, "BFAST: An alignment tool for large scale genome resequencing," *PLOS One*, vol. 4, no. 11, p. e7767, 2009.

[6] F. Krueger and S. R. Andrews, "Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications," *Bioinformatics*, vol. 27, no. 11, pp. 1571–1572, 2011.

[7] B. S. Pedersen, K. Eyring, S. De, I. V. Yang, and D. A. Schwartz, "Fast and accurate alignment of long bisulfite-seq reads," *arXiv preprint arXiv:1401.1129*, 2014.

[8] J. Porter and L. Zhang, "BisPin and BFAST-Gap: Mapping bisulfite-treated reads," *bioRxiv*, p. 26, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/03/19/284596

[9] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, 2012, pp. 2951–2959.

[10] L. S. Zou, M. R. Erdos, D. L. Taylor, P. S. Chines, A. Varshney, S. C. Parker, F. S. Collins, J. P. Didion *et al.*, "BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues," *bioRxiv*, p. 207506, 2018.

[11] Y. Wang, T. Liu, D. Xu, H. Shi, C. Zhang, Y.-Y. Mo, and Z. Wang, "Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks," *Scientific Reports*, vol. 6, p. 19598, 2016.

[12] J. He, M.-a. Sun, Z. Wang, Q. Wang, Q. Li, and H. Xie, "Characterization and machine learning prediction of allele-specific DNA methylation," *Genomics*, vol. 106, no. 6, pp. 331–339, 2015.

[13] A. Vidaki, D. Ballard, A. Aliferi, T. H. Miller, L. P. Barron *et al.*, "DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing," *Forensic Science International: Genetics*, vol. 28, pp. 225–236, 2017.

[14] J. Naue, H. C. Hoefsloot, O. R. Mook, L. Rijlaarsdam-Hoekstra, M. C. van der Zwalm, P. Henneman, A. D. Kloosterman, and P. J. Verschure, "Chronological age prediction based on dna methylation: Massive parallel sequencing and random forest regression," *Forensic Science International: Genetics*, vol. 31, pp. 19–28, 2017.

[15] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, p. 321, 2015.

[16] J. Porter, M.-a. Sun, H. Xie, and L. Zhang, "Investigating bisulfite short-read mapping failure with hairpin bisulfite sequencing data," *BMC Genomics*, vol. 16, no. 11, p. S2, 2015.

[17] R. Leinonen, H. Sugawara, M. Shumway, and I. N. S. D. Collaboration, "The sequence read archive," *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D19–D21, 2010.

[18] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* University of Illinois Press, 1949.

[19] V. Phan, S. Gao, Q. Tran, and N. S. Vo, "How genome complexity can explain the difficulty of aligning reads to genomes," *BMC Bioinformatics*, vol. 16, no. 17, p. S3, 2015.

[20] J. Porter and L. Zhang, "InfoTrim: A DNA read quality trimmer using entropy," in *Computational Advances in Bio and Medical Sciences (ICCABS), 2017 IEEE 7th International Conference on*. IEEE, 2017, pp. 1–2.

[21] J.-Q. Lim, C. Tennakoon, G. Li, E. Wong, Y. Ruan, C.-L. Wei, and W.-K. Sung, "BatMeth: Improved mapper for bisulfite sequencing reads on DNA methylation," *Genome Biology*, vol. 13, no. 10, p. R82, 2012.