Machine learning algorithms in the prediction of conflicts in clinical classification of genetic variants

Kirill Musin
Samara National Research University Samara, Russia
kmusin07@gmail.com

Andrey Gaidel
Samara National Research University;
Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS
Samara, Russia
andrey.gaidel@gmail.com

***Abstract*—The clinical classification of a person's genetic variant can lead to conflicting classifications. The presence of conflicts is determined manually by laboratory methods. If there is a conflict, then there is a difficulty in interpreting the result. In this work, with the help of machine learning algorithms, it was possible to train the neural network to predict conflicts with an accuracy of 77%, and also to determine which parameters are most important in classification.**

***Keywords*—machine learning, classification of conflicts, prediction, biology, medicine, variations of the human genome**

I. Introduction

The constant discoveries in biology provide us with a huge amount of data for analysis, studying this data we can find more and more patterns and relationships between the various characteristics of living organisms, which leads us to a greater understanding of how the world works and how we can improve it. One of the studies that triggered a push in the development of genetics was done by the biologist Gregor Mendel. His works contain information on the established relationship between the presence of certain genes and various morphological and physiological characteristics of individuals, as well as on such a key property of organisms that the genetic code is able to be transmitted hereditarily, with preservation of signs from parents to descendants [1].

Starting with the study of simple organisms, research has reached the human genome. One of the directions in this environment is a genome-wide search for associations related to the study of associations between genomic variants and phenotypic characters. The main goal is to predict a predisposition to a disease by identifying genetic risk options, which is based on a comparison of the alleles of healthy people and people with diseases [2]. Thanks to such studies, genetic variants were obtained that affect the risk of complex cardiovascular diseases, autoimmune diseases, and cancer [3, 4, 5].

The studies mentioned above are carried out empirically by specialists in this field. T he d evelopment o f information technologies allows us to simplify and automate the same type of work, so using machine learning methods, the researchers in [6] built a model in which the relationship between many different single nucleotide polymorphisms and complex human diseases is determined. Researchers, using various methods of machine learning, received several qualitative phenotypes that affect the interaction of genes. Also, based on models of ensemble methods, the most important variables affecting the genome were identified. But the authors led to the fact that there were some limitations that allow us to consider the use of machine learning methods as an addition to existing laboratory tests to identify the basis of complex genetic diseases [6].

A study was conducted in [7], during which it was proved that machine learning provides an additional look at the analysis of multidimensional genetic sets in comparison with standard approaches to testing statistical associations. It has also been shown that approaches to multifactorial modeling allow a better understanding of the genetic characteristics of diseases, as was the case with atherosclerosis, coronary heart disease, and elevated lipid levels. But the authors caution future researchers that when using machine learning methods in genetics, you need to carefully make sure that there is no risk of retraining the model, which can lead to overly optimistic forecasting results. The authors argue that in order to solve the problem of predicting the risk of a genetic disease, it is necessary to carry out effective regularization and strict validation of the model [7].

As already mentioned, various genetic variations entail a predisposition or development of various diseases. Modern studies show that not everything can be established unambiguously, speaking about genetics, since situations are possible when a patient has a disease, but his genetic set does not correspond to this disease, or vice versa. Such situations are called conflict situations and, if they arise, additional laboratory tests are required that require large investments to determine if this combination can really be associated with the disease. Data on such studies and their results are presented in the public domain on the ClinVar portal. Due to the emergence of a number of difficulties for solving the genetic conflict by the laboratory method, this paper presents a different approach based on the work of machine learning methods and data generated on the basis of clinical results [8].

In this paper, using various methods of generating features: Feature Hasher, One Hot Encode, Label Encoder, as well as custom data, we prepare data for further processing using machine learning methods such as Random Forest and Gradient Boosting. A detailed analysis is also carried out to identify the best model for solving the classification problem of genetic conflict.

II. Data preparation for model training

A. Feature engineering of the current issue

Machine learning methods can work only with data presented in numerical form. Also, based on work [9], it is necessary to form a small subset of features from a large set of source data that would be most effective for solving the problem. It is also necessary to consider that due to the fact that there are more patients without genetic conflict than without them, hence, machine learning models will also tend to this result.

Non-numerical features were processed. The number of unique values in them influenced the choice of method for processing.

The PolyPhen (Polymorphism Phenotyping) characterization consists of several descriptions of the possible effects of amino acid substitution on the structure and function of the human protein. For this data set, the Label Encoder method was applied, which assigns a certain number to each state. Such a characteristic as EXON (Exon), a part of the gene encoding amino acids, was presented in the source data as a part of exons of their total number in the body, which required additional processing using methods for working with strings from the Python library.

For characteristics such as cDNA position (position of the gene pair in the sequence of additional DNA), CDS position (position of the base pair of genes in the coding region), Protein position (position of the amino acid in the protein), the data presented as ranges of positions; therefore, their median values expressed scalar.

Characteristics: REF (comparison allele), ALT (alternative allele), CHROM (chromosome variant), Allele (allele), Consequence (consequence type) contain a number of different values from 24 to 866. Because of this, we can consider the applicable encoding method for these values The Feature hasher method, which vectorizes features into a certain number of columns that can represented as elements of Boolean algebra.

Such characteristics as: CLNVC (variant type), IMPACT (impact modifier for the kind of consequence), BIOTYPE (biotype) have a relatively small set of different values. Therefore, the One Hot Encode method is well applicable for them, creating a vector for each characteristic value.

Studies also conducted on the remaining columns; in some of them, there were no values, which would only complicate the work of the Random forest method.

In some, like CLNHGVS (a characteristic containing a description of the level of genome location), all values were unique, it follows that the correlation between them is zero. The establishment of relationships that play a role for the classifier is impossible. Therefore, part of the information has been removed.

III. Selected machine learning methods

This forecasting problem can be solved using binary classification methods [10]. The ensemble methods applied to this task are: Random forest, whose ensemble consists of simple models called the Decision tree (Decision tree), as well as Gradient boosting, which has an exceptionally different way of interacting with the base models.

A. Description of the mathematical model of the selected methods

The decision tree described as follows: let the training vectors $x_i \in R^n, i = 1, \ldots, l$, and the label vector $y \in R^l$ given: the decision tree divides the space recursively so that the samples with the same labels grouped. Let the data at node m represented by Q. For each candidate, the separation $\theta = (j, t_m)$, consisting of the characteristic j and the threshold t_m, divides the data into subsets $Q_{left}(\theta)$ and $Q_{right}(\theta)$, where $Q_{left}(\theta) = (x, y)|x_j \leq t_m, Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$. Then, the function $G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$ is calculated, where $H(\theta)$ is the measure of entropy given by the formula: $H(X_m) = \sum_k p_{mk} log(p_{mk})$. Next, the parameters selected according to the following criterion: $\theta^* = argmin_\theta G(Q, \theta)$. The subsets $Q_{left} = (\theta^*)$ and $Q_{right} = (\theta^*)$ are determined until the maximum available depth is reached:$N_m < min_{samples}$ or $N_m = 1$ [11].

B. Difference between Random Forest and Gradient Boosting

The interaction of decision trees in the Random Forest algorithm carried out using the bagging approach - the creation of independent models for assessment, and then the averaging of their forecasts using the following formula: $S_l = \frac{1}{l} \sum_{l=1}^{L} w_l$, where L is the number of independent base models, and w_l is the received dataset by each model [12]. This approach leads to less dispersion.

In the Gradient boosting method, interactions between decision trees carried out according to the principle of boosting, based on the fact that the family of models is combined to create the strongest of the basic ones. Several weak models are adaptively selected, and based on their results. A stronger value is attached to those objects in the dataset that were poorly processed by previous models, thus reducing the bias of the estimate even with a decrease in the spread [13].

Gradient boosting considers additive models of the following form:$F(x) = \sum_{m=1}^{M} \lambda_m h_m(x)$, where $h_m()$ are the basic models, and λ_m is the step size calculated by the one-dimensional optimization process. This method, like the Random Forest, uses the decision tree as a simple model. Thanks to such features of the "tree" as processing mixed-type data and modeling complex functions, it is ideal for optimizing the step. GB constructs the additive model in a "greedy" way: $F_m(x) = F_{m-1}(x) + \lambda_m h_m(x)$, where the recently added tree h_m tries to minimize the loss of L, given the previous ensemble F_{m-1}: $h_m = \underset{h}{argmin} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x_i))$ [14].

IV. The result of training models

For the search of the best models of Random Forest and Gradient boosting, it was necessary to identify the parameters that give the best result. As the metrics, basic ones were

used, such as: Precision, Recall, F-score. These metrics are calculated based on the following criteria: True Positives (TP) - the correctly predicted positive value of the source class; It is worth noting that for binary classification, the positive result is one, and the negative is zero; True Negatives (TN) - the correctly predicted negative value of the source class, False Positives (FP) - the case when the result in the original class is represented by negation, while the classifier returned a positive result; False Negatives (FN) - the case is the opposite of FP. Precision is the ratio of correctly predicted values to the total number of attempts to give a positive result, presented by the formula: $\frac{TP}{TP+FP}$. Recall is the ratio of correctly predicted values to their total number: $\frac{TP}{TP+FN}$. F-score is the most accurate measure, which contains the two scores listed above, given by the formula: $2\frac{Recall\ x\ Precision}{Recall+Precision}$ [15].

Consider how the results of the models depend on the number of trees. For Random Forest, the result of the dependence of F-score, Precision, Recall on the number of trees in the ensemble is clearly shown in Figure 1. As you can see, the Precision value actively increases with the number of trees. Still, after 64, the result remains practically unchanged in the Recall, and The F-score also has such a tendency with a large number of trees. However, the graph also shows that they take the most excellent value with five trees, in which case the Precision, Recall, F-score measures take the following values, respectively: 0.46, 0.37, 0.41. But, unfortunately, a model with so many trees is not suitable for correction using the predict proba method, since the essence of this method is that you can control the adoption of the tree's voice by increasing the weights, then with a small number of trees the result of such a change is rather weak. Therefore, it will be entirely objective to take a model with 128 trees.

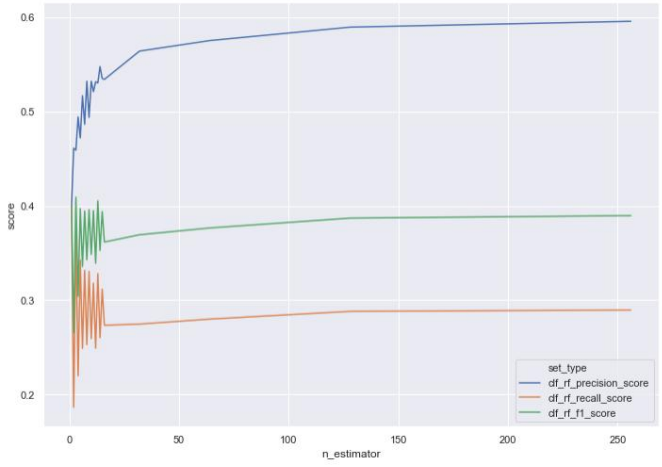

Fig. 1. The dependence of the result of metrics on the number of trees in Random Forest.

The result of such a model recorded in table 1; the table also contains the result after applying predict proba.

The same study done for Gradient boosting, the metrics for a different number of trees shown in Figure 2. Because each of the following trees of this method is trained based on previous

TABLE I. RANDOM FOREST RESULTS WITH A MODEL OF 128 ESTIMATORS

Without modification	Precision	Recall	F-score
0	0.79	0.93	0.85
1	0.57	0.28	0.38
With modification			
0	0.9	0.62	0.73
1	0.41	0.8	0.55

results, it found experimentally that less than 14 trees are not enough to predict results. But after 14, an apparent increase in the classification accuracy is visible, so it would be reasonable to take the one that has the largest number of trees as the main model and modify it by changing the weights. Still, practical measurements have shown that for this task, the classifier models based on Gradient boosting with the number trees 128, 256, 512 have the same result after applying the predict proba method. In this case, for a more objective comparison with the Random Forest method, we take a model with 128 trees.

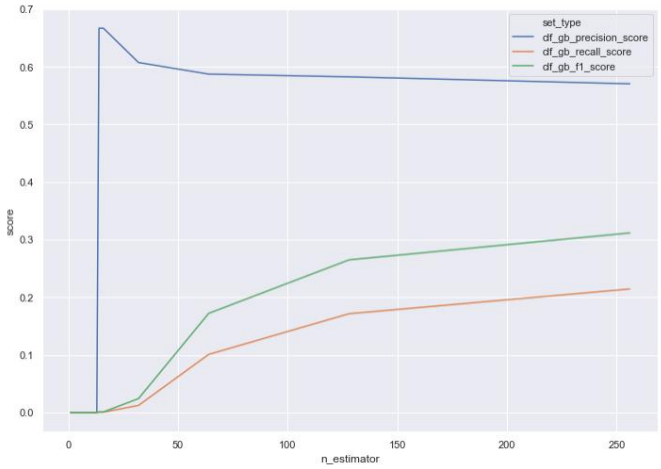

Fig. 2. The dependence of the result of metrics on the number of trees in Gradient Boosting.

The result of the best classifier m odel b ased o n Gradient boosting presented in Table 2. Comparing the results from

TABLE II. GRADIENT BOOSTING RESULTS WITH A MODEL OF 128 ESTIMATORS

Without modification	Precision	Recall	F-score
0	0.79	0.93	0.85
1	0.56	0.28	0.37
With modification			
0	0.87	0.75	0.8
1	0.47	0.66	0.55

Tables 1 and 2, it becomes clear that the results of the metrics before modification are identical for the two classifiers; after applying the predict proba method, the Precision and Recall metrics are different, but for a positive case, the F-score metric is the same. For an additional comparison of methods, to find the optimal one, Table 3 shows the time during which the

classifiers a re t rained a nd a lso p redict t he v alues o n t he test sample.

TABLE III. GRADIENT BOOSTING AND RANDOM FOREST TIME COMPARISON RESULTS WITH THE MODELS OF 128 ESTIMATORS

	Training time, sec.	Classification time, sec.
Random Forest	11.0	0.55
Gradient Boosting	8.36	0.05

The training time for the two methods turned out to be comparable, while the classification time for the model based on Gradient Boosting is ahead of the model based on Random Forest by order of magnitude. Thus, the Gradient Boosting classifier is most preferred for determining whether a patient has a genetic conflict.

Also, Figure 3 shows a graph of the ROC-curve for two models, which allows one to evaluate the quality of the binary classification [11].

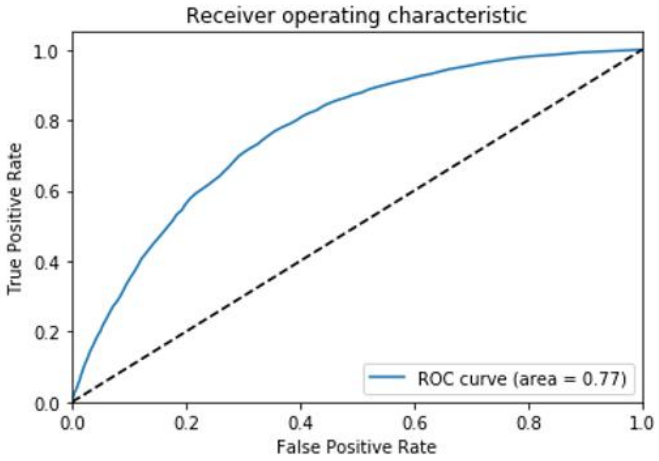

Fig. 3. ROC-curve for determining the quality of binary classification.

Based on Figure 2, the reading area is estimated at 0.77, the quality of the classifier i s d etermined b y h ow m uch this indicator is high.

V. SUMMARY

In this paper, we studied the effectiveness of machine learning algorithms for the task of establishing genetic conflict in clinical analysis. It also found that the Gradient Boosting method is preferable for this task, since the trained model copes with forecasting ten times faster than the model based on Random Forest. The obtained accuracy result of 79% shows that the use of machine methods for the tasks of genetic classification a nd c onflict re solution is po ssible, de spite the difficulty i n i nterpreting t he i nitial p arameters f or humans. The introduction of trained models will reduce the cost of additional medical research, and will also predict the need for screening for patients with suspected illness.

ACKNOWLEDGMENT

The work was partially funded by the Russian Foundation for Basic Research under grants No. 19-29-01235 and 19-29-01135 (theoretical results) and the RF Ministry of Science and Higher Education within the government project of the FSRC Crystallography and Photonics RAS under grant No. 007-GZ/Ch3363/26 (numerical calculations).

REFERENCES

[1] G. Mendal, "Experiments in plant hybridization," Cosimo Classics, p. 52, 2008.
[2] W.S. Bush and J.H. Moore, "Genome-wide association studies," Public Library of Science for Computational Biology, vol. 8, 1002822, 2012. DOI: 10.1371/journal.pcbi.1002822.
[3] K.L. Mohlke, M. Boehnke and G.R. Abecasis, "Metabolic and cardio-vascular traits: an abundance of recently identified common genetic variants," Hum. Mol. Genet., vol. 17, pp. 102-108, 2008. DOI: 10.1093/hmg/ddn275.
[4] G. Lettre and J.D. Rioux, "Autoimmune diseases: insights from genome-wide association studies," Hum. Mol. Genet., vol. 17, pp. 116-121, 2008. DOI: 10.1093/hmg/ddn246.
[5] D.F. Easton and R.A. Eeles, "Genome-wide association studies in cancer," Hum. Mol. Genet., vol. 17, pp. 109-115, 2008. DOI: 10.1093/hmg/ddn287.
[6] S. Szymcak, J.M. Biernacka, H.J. Cordell, O. Gonzales-Recio, I.R. Konig, H. Zhang and Y.V. Sun, "Machine learning in genome-wide association studies," Genet Epidemiol, vol. 33, pp. 51-57, 2009. DOI: 10.1002/gepi.20473.
[7] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti and T. Aittokallio, "Regularized Machine Learning in the Genetic Prediction of Complex Traits," PLoS Genet, vol. 11, 1004754, 2014. DOI: 10.1371/jour-nal.pgen.1004754.
[8] ClinVar [Online]. URLhttps://www.ncbi.nlm.nih.gov/clinvar/.
[9] A.V. Gaidel, "Matched polynomial features for the analysis of grayscale biomedical images," Computer Optics, vol. 40, no. 2, pp. 232-239, 2016. DOI: 10.18287/2412-6179-2016-40-2-232-239.
[10] "Classification," MachineLearning.ru [Online]. URL: http://www.machinelearning.ru/wiki/index.php.
[11] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, "Classification and regression trees," International Biometric Society, 1984. DOI: 10.2307/2530946.
[12] "Desicion trees," Scikitlearn [Online]. URL: https://scikit-learn.org/stable/modules/tree.html.
[13] G. Louppe, P. Geurts, P.A Flach, T. De Bie and N. Cristianini, "Ensembles on Random Patches," Machine Learning and Knowledge Discovery in Databases, 2012. DOI: 10.1007/978-3-642-33460-3-28.
[14] "Ensemble methods," Scikitlearn [Online]. URL: https://scikit-learn.org/stable/modules/model-evaluation.html.
[15] "Metrics and scoring: quantifying the quality of predictions," Scikitlearn [Online]. URL: https://scikit-learn.org/stable/modules/model-evaluation.html.