 Future Technologies and Innovations (FTI) Proceedings: 4th international conference on computer applications and information security (iccais'2021) / March 19 / 2021/ Tunisia: https://fti-tn.net/iccais-2021-list-of-papers

Analysis of Genetic Variants Conflict Classification using various Machine Learning Algorithms

Fathima Mirza
Department of CSE
BRAC University
Dhaka, Bangladesh
fathima.mirza@g.bracu.ac.bd

***Abstract*— The ClinVar dataset contains information of human genetic variants which forms a two-class classification problem. This dataset was first reduced using Principal Component Analysis for dimension reduction and then various Machine Learning Algorithms were applied to a random sample of the dimension reduced dataset to see which Algorithm performed best. Support Vector Machine Classifier returned the highest accuracy score.**

Keywords—Logistic Regression, KNN, Nearest Centroid, Decision Tree, Random Forest, AdaBoost, Gradient Boost, SVM, Classification, Genetic variant

Introduction

ClinVar is a public database that collects data about human genetic data and its association to human health. This dataset is publicly available on Kaggle [1]. One of the information contained in the dataset is the annotation about genetic variants. These variants are (usually manually) classified by clinical laboratories on a categorical spectrum ranging from benign, likely benign, uncertain significance, likely pathogenic, and pathogenic. Variants that have conflicting classifications (from laboratory to laboratory) can cause confusion when clinicians or researchers try to interpret whether the variant has an impact on the disease of a given patient [1]. The aim of this paper is to make a prediction about the conflicting classes present in the ClinVar dataset. This forms a binary classification problem where 0 in the column CLASS represents consistent classification and 1 in the column CLASS represents conflicting classification [1]. This can be further explained as follows: Two patients visit two different labs to test a particular genetic variant. If the particular variant produces the same or similar classification by both the geneticists at both the labs, then it is said to be consistent classification, otherwise, it falls under the conflicting classifications. A class is said to be conflicting if the same variant consists of two different categories of results, and is said to be consistent if the same variant consists of the same categories of result. The categories are as follows: a) Likely benign or benign b) VUS c) Likely pathogenic or pathogenic [1]. It is important to note that if a variant produces a result of benign and likely benign, that variant will be classified as consistent because it falls under the same category. This is diagrammatically explained in Fig 1 and Fig 2 [1].

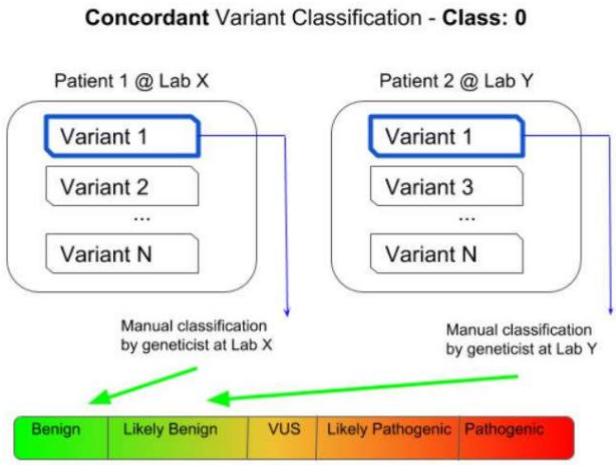

Fig 1: Concordant or Consistent Classification CLASS 0

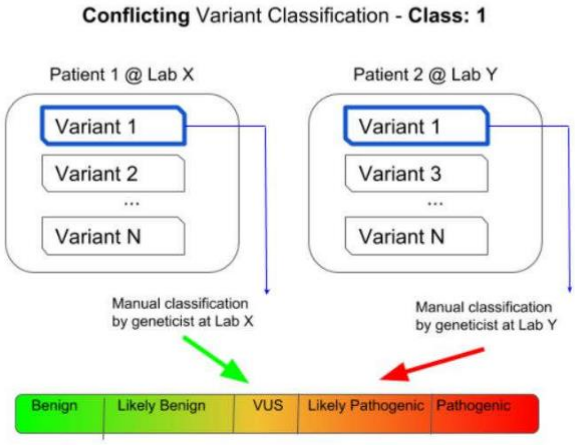

Fig 2: Conflicting Variant Classification CLASS 1

I. Related Work

ClinVar at the National Center for Biotechnology Information (NCBI) is a freely available archive for interpretations of clinical significance of variants for reported conditions which contains germline and somatic variants of any size, type or genomic location. Interpretations are submitted by various labourites which are reviewed by ClinVar staff with data types such as HGVS (Human Genome Variation Society) expressions. Clinical significance is calculated for the aggregate record, indicating consensus or conflict in the submitted interpretations [12]. Clinotator is a fast and lightweight tool to extract important aspects of criteria-based clinical assertions; it uses that information to generate several metrics to assess the strength

and consistency of the evidence supporting the variant clinical significance. Clinical assertions are weighted by significance type, age of submission and submitter expertise category to filter outdated or incomplete assertions that otherwise confound interpretation. Clinotator provides efficient, systematic prioritization of discordant variants in need of reclassification [13]. Automatic Variant evidence DAtabase (AVADA) is a novel machine learning tool that uses natural language processing to automatically identify pathogenic genetic variant evidence in full-text primary literature about monogenic disease and convert it to genomic coordinates [14]. A deep learning model to accurately predict locus-specific signals from four epigenetic assays using only DNA sequence as input is developed. Given the predicted epigenetic signal from DNA sequence for the reference and alternative alleles at a given locus, a score of the predicted epigenetic consequences for 438 million variants observed in previous sequencing projects was generated. These impact scores are assay-specific, are predictive of allele-specific transcription factor binding and are enriched for variants associated with gene expression and disease risk. Nucleotide-level functional consequence scores for non-coding variants can refine the mechanism of known functional variants, identify novel risk variants and prioritize downstream experiments [15]. It was demonstrated that the neural network model AIVAR (Artificial Intelligent VARiant classifier) was highly comparable to human experts on multiple verified datasets. Although highly accurate on known variants, AIVAR together with CADD and PhyloP showed non-significant concordance with SGE function scores. Moreover, our results indicated that neural network model trained from functional assay data may not produce accurate prediction on known variants [16].

II. Proposed Method

The purpose of this project or paper is to predict the conflicting classifications in the genetic variants. To make the predictions, the dataset was first prepared to ensure that the best result will be generated. The missing values were filled with 0's. The dataset has 45 features. To make certain that only the most relevant features were taken into action when it came to implementing the algorithms, Principal Component Analysis (PCA) which is a technique to demonstrate the correlations between the features in a dataset and can be used for reducing dimensionality was applied to select the two most pertinent features in the dataset. The dataset was randomly sampled and 1000 instances were selected. This was done for two main reasons a) to reduce the time needed to run the analysis and b) because it helps out cancel the effect of unobserved data and reduce biases [2]. Out of the 1000 instances, 70% was used to train the data and 30% was used as test data. Once the data was prepared, Logistic Regression, K-Nearest Neighbours, Nearest Centroid, Decision Tree, AdaBoost, Gradient Boost, Random Forest and Support Vector Machine algorithms were applied on the training dataset to train the models and was used on the test dataset to predict the classification. Confusion matrix which can be used to describe how well the algorithms perform, was computed for all of the algorithms. From the confusion matrix, the accuracy score of all the matrix was calculated as follows:

$$\text{Accuracy Score} = (TP + TN)/\{TP + TN + FP + FN) \quad (1)$$

where TP stands for True Positive which is when the predicted 1 matches the actual 1, TN stands for True Negative which is when the predicted 0 matches the actual 0, FP stands for False Positive also known as a Type I error, which is when the predicted is 1 but the actual is 0, and False Negative also known as a Type II error, which is when the predicted is 0 but the actual is 1.

III. Comparative Study

Various sampling values were tested as well as various values for testing and training data. In addition, the algorithms were hyper tuned, and the results were compared with the default values. For this particular problem, the default settings gave the most optimum results. Therefore after rigorous trials, a sampling value of 1000, a testing value of 30% of the dataset and default settings for all the algorithms was chosen for this project or paper. Logistic Regression is a Machine Learning algorithm which is used for classification problems used to assign observations to a discrete set of classes. It is a predictive analysis algorithm and based on the concept of probability, which uses a sigmoid function [3]. The confusion matrix for Logistic Regression for this particular conflict classification problem is [[137, 94], [39, 30]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in equation (1), we get an accuracy score of 0.556667. The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised (relies on labelled input data to learn a function that produces an appropriate output when given new unlabelled data) machine learning algorithm that can be used to solve classification

[4]. The confusion matrix for KNN for this particular conflict classification problem is [[211, 20], [62, 7]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in equation (1), we get an accuracy score of 0.726667. Nearest centroid classifier or nearest prototype classifier is a classification model that assigns to observations the label of the class of training samples whose mean (centroid) is closest to the observation
[5]. The confusion matrix for nearest centroid for this particular conflict classification problem is [[138, 93], [39, 30]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in "(1)", we get an accuracy score of 0.560000. Decision Tree uses a tree-like model of decisions to derive a classification result [6]. The confusion matrix for Decision Tree for this particular conflict classification problem is [[180, 51], [49, 20]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in "(1)", we get an accuracy score of 0.666667. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [7]. The confusion matrix for Random Forest for this particular conflict classification problem is [[208, 23], [62, 7]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in "(1)", we get an accuracy score of 0.716667. Random Forest performs better than decision trees because it solves the overfitting problem that exists in Decision Trees. AdaBoost is one of the first boosting algorithms to be adapted in solving practices that

combines multiple "weak classifiers" into a single "strong classifier" [8]. The confusion matrix for AdaBoost for this particular conflict classification problem is [[226, 5], [66, 3]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in "(1)", we get an accuracy score of 0.763333. Gradient Boosting trains many models in a gradual, additive and sequential manner by using gradients in the loss function whereas AdaBoost identifies the shortcomings by using high weight data points [9]. The confusion matrix for Gradient Boosting for this particular conflict classification problem is [[221, 10], [62, 7]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in "(1)", we get an accuracy score of 0.760000. Support Vector Machine (SVM) is a machine learning algorithm that finds a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points [10]. The confusion matrix for SVM for this particular conflict classification problem is [[231, 0], [69, 0]] where the format is as follows [[TP, FP],[FN,TN]]. Using the values from the confusion matrix and substituting them in "(1)", we get an accuracy score of 0.770000. SVM under default settings where the Regularization Parameter (C) is set to 1 and the kernel is set to rbf performs most superiorly compared to all the other classification algorithms. SVM performs best for binary classifications is more robust i.e. due to optimal margin gap between separating hyper planes hence predicts more accurately than the other algorithms [11].

The above mentioned accuracy scores are demonstrated in Table 1 and Fig 3 as a bar plot in the ascending order of their performance.

TABLE 1: Accuracy of Models

Model	***Accuracy***
Logistic Regression	0.5566667
Nearest Centroid	0.56
Decision Tree	0.6666667
Random Forest	0.7166667
KNN	0.7266667
Gradient Boost	0.76
AdaBoost	0.7633333
SVC	0.77

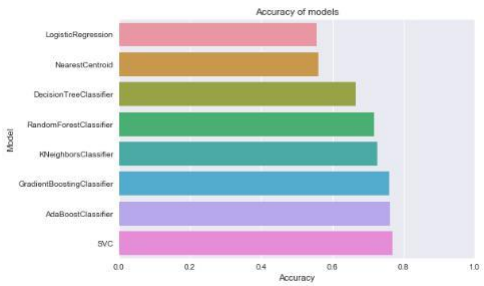

Fig 3: Accuracy of models

IV. Conclusion

This paper focuses on several classification machine learning algorithms to decide whether a variant is of conflicting or consistent classification. The algorithms that were discussed in detail in this paper are Logistic Regression, KNN, Nearest Centroid, Decision Tree, Random Forest, AdaBoost, Gradient Boost and SVM. SVM performs the best for this dataset and problem specification and AdaBoost ranks a close second. The performance could perhaps be further enhanced by carrying out intensive data preprocessing techniques. This shall be attempted as a future work for this dataset.

Acknowledgment

I would like to thank Dr. Md. Golam Rabiul Alam from BRAC University and G. M. Shahariar from Ahsanullah University of Science and Technology who provided insight and expertise that greatly assisted the project.

References

[1] https://www.kaggle.com/kevinarvai/clinvar-conflicting
[2] https://stats.stackexchange.com/questions/41770/random-sampling-real-data-is-so-important-why
[3] https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
[4] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
[5] https://en.wikipedia.org/wiki/Nearest_centroid_classifier
[6] https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
[7] https://en.wikipedia.org/wiki/Random_forest
[8] https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe
[9] https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab
[10] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
[11] https://www.researchgate.net/post/Why_Support_Vector_MachineSV M-Best_Classifier
[12] Landrum, M. J. (2015). ClinVar: Public archive of interpretations of clinically relevant variants.
[13] III, R. R. (2018). Clinotator: analyzing ClinVar variation reports to prioritize reclassification efforts.
[14] Birgmeier, J. (2019). AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature. *Genetics in Medicine.*
[15] Hoffman, G. E. (2019). Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. *Nucleic Acids Research.*
[16] Luo, J., Zhou, T., You, X., Zi, Y., Li, X., Wu, Y., & Zhaoji Lan. (2019). Assessing concordance among human, in silico predictions and functional assays on genetic variant classification. *Bioinformatics.*