International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Improving Binary Classification Accuracy Using Stacked Ensemble of Diverse Models

Dr. Vidya Chitre[1], Sruthi Nair[2]

[1]Assistant Professor, Vidyalankar Institute of Technology, Mumbai, India
Email: *vidya.chitre[at]vit.edu.in*

[2]Master of Engineering, Vidyalankar Institute of Technology, Mumbai, India
Email: *shruthi.nair[at]vit.edu.in*

Abstract: ***Every Machine Learning algorithm has some advantages and disadvantages of it's own, but all have the common error of high - dimensional feature set overfitting the training data. This causes depletion in performance by driving algorithm into generalization error. One of the Ensemble Learning methods called Stacking or Stacked Generalization can solve this problem. In this paper we carry out binary classification using Stacked Generalization on high dimensional Polycystic Ovary Syndrome dataset and demonstrate that model generalizes and metrics such as accuracy improve substantially. There are several other metrics which in my opinion provide a glaring pg57 with Receiver Operating Characteristic Curve which provides evidence of incorrectness.***

Keywords: Ensemble Learning, Generalizing Error, Stacked Generalization

1. Introduction

No matter how effective a Machine Learning technique is, drawing inference from high dimensional data remains very challenging. For a machine learning model, inference that cannot be derived results in loss of information, which is termed error. Deep learning techniques provide loss optimizers which help reduce the error, but machine learning absolutely needs optimizers for error generalization. If you look closely, there are a plethora of algorithms with both parametric and non - parametric learning technique branches available for machine learning. The learning methods focus on how well the model fits the data regardless of its dimensional space. Focussing on classes, we will concentrate on binary classification as that's the type of classification this paper is based on. Logistic Regression is one of the most widely adopted algorithms for binary classification. Logistic Regression comes under the family of supervised learning, parametric and uses a logit function to give 2 distinct classes. Even after trying to tune hyperparameters of an algorithm, there are always going to be algorithms that are optimal for other aspects of data. Extensions of linear models can include K - Nearest Neighbors, Support Vector Machines and many more.

Our selection criteria for the data was not constrained as we needed a data set with a very high amount of dimensions. The data we chose for this work is Polycystic Ovary Syndrome Classification which is specifically a binary classification problem based on the features gives the presence of symptom. The data is very high dimensional in terms of features and there are numerous categorical features. In order to leverage any machine learning algorithms' performance, categorical parameters are difficult to the ground as the basis of it when converted into numbers depends upon its goal. Performance for the algorithms starts at diagonstic performance hurdles. There is highly characteristic dependent conversion classification needed for such variables. There is need for transformation of such variables in rank oriented versus occurrence oriented formats and we analyzed many facets. If we analyze polycystic ovary syndrome under the machine learning veneer, the problem is tackled using logistic regression, bagging ensemble, discriminant analysis, and boosting ensemble methods are the focus in these papers. This document addresses primarily methodologies aimed at generalizing the errors of other models and secondarily looks into application of the polycystic ovary syndrome data through the lens of the methodology offered in this paper.

2. Methodology

In this paper, this part outlines the step we take in solving the problem. To address the issue of generalization error, we implement the Stacking Ensemble Method which is also referred to as Stacked Generalization. We will walk you through all the classifiers which form the base of the stack before we delve into explaining the stacked model.

a) Logistic Regression

This is going to be one of the algorithms that contribute to the stack. Logistic regression is a form of supervised learning which involves estimating a linear equation. The logistic regression computes fixed parameters and applies them to compute a prediction equation which is analogous to linear regression. For example, the prediction function for computing a single feature can be expressed as:

$$y\hat{}=y0+y1xiy\hat{}=y0+y1xi$$

The equation above aims to fit a straight line to a set of data points. However, in order to perform classification using logistic regression, the line must be translated through the logit function, which will provide clear delineation between the classes. The function for logit is given as:

$$\sigma(z) = 11+e-z\sigma(z) = 1+e-z1$$

Here, is termed as logit, which is also called Sigmoid function while z is the linear function.

b) Support Vector Machine

Support Vector Machine (SVM) is one of the supervised learning algorithms in which a non - parametric approach is taken to estimate the hyperplane function over data points. A hyperplane is a line that is drawn within a data set basin such that classes can be separated stratum by stratum. Hyperplane is a single line and get as close to as maximum margin possible to the points. The elements which 'touch' the margin are called Support Vectors. For hyperplane calculation, a vector normal and one offset point need to be determined. This can be backed up in equation format as:

$$w \cdot x+b=0 \; w \cdot x+b=0$$

Hypothesis or the decision boundary for classifying data can also be described using a normal vector w, which is generated automatically by machine learning algorithms during the process of training. b is called the offset value.

c) Multi Layer Perceptron

An acronym of Deep Learning terms which include Multi Layer Networks is Multi Layer Perceptron (MLP). It has its roots as Perceptron. What Perceptron does is that it has weights and biases that are set up to begin with and these impact how representations in the data are offloaded. With these biases and weights, an activation function has to be added so as to preserve the features of the data. Among the family of Non - Linearity Activation functions is Rectified Linear Unit (ReLU).

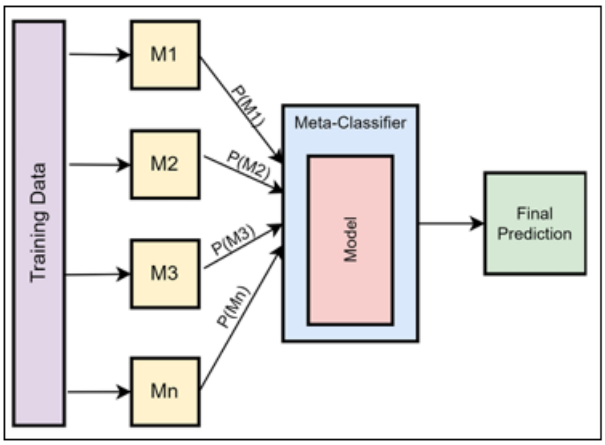
d) Random Forest

A non - parametric supervised bagging ensemble learning method is called random forest. It has weak learners and was built to enhance the decision tree which has a high variance problem because of too many features. The random forest is also a bootstrap aggregating method as it involves combining a set of trees and using their collective result.

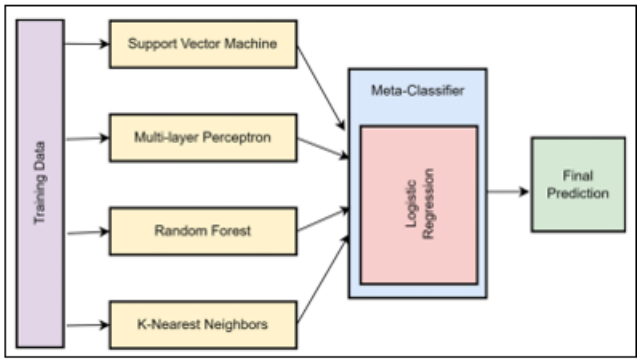
e) K - Nearest Neighbors

The method of classification known as K - Nearest Neighbors (KNN) is based on supervised non - parametric distance learning. The algorithm takes into account distance learning equations for class estimation based on the new data points available for prediction. Classes with the highest predicted probabilities are taken as the final output using majority votes.
F. Stacked Generalization

This is where the critical part of the paper rests, the stacked generalization model, and we will explain it along with all the previous discussions. There are two phases in stacked generalization. In the first phase, a diversified ensemble of models is built through training, and their predictions are taken. In the second phase, the model is trained using the predictions made during phase one and the actual class labels of the prediction. This latter phase is referred to as meta - classifier, and it only serves to provide the class label for another piece of data that was already evaluated by the first phase. The depiction can be done in the Fig 1 given below.



The models in the first stage of Figure 1 operate independently of one another. They can be non - parametric or parametric, might number into the hundreds or even more, and could be of the same type differing by model hyperparameters. The outcomes of every model are generated and then, handed over to the meta - classifier block. This meta - classifier is also a model, which could be either parametric or non - parametric, and serves as a dependent model which takes in the output of the first stage in the figure as input. It later produces the final output to the given input. This approach serves to combine and generalize the predictions of the models and helps to alleviate the shortcomings posed when models are used independently. In implementing example, the models we employed for were support vector machine, random forest, multi - layer perceptron and k - nearest neighbors. We implemented a linear model and logistic regression for the meta - classifier, which provided the final outcome alongside the other models. And this is how we achieve our final output.



In our stacking model, shown in Figure 2, we utilize four base models in the first stage, with a Logistic Regression meta - classifier for generalization. The training parameters were the same for all classifiers. SVM defaults were adopted with an RBF kernel and gamma scaling; MLP's had ReLU with alpha set to 0.1, 1000 hidden units. Random Formest had 500 estimators with Gini criterion, max depth of 10, min leaf sample of 0.005. KNN was set with k equaling 5 and uniform weights. For the stacking Logistic Regression classifier with cross validation set to 5 folds, Logistic Regression was trained with default hyper parameters.

Algorithm	F0.5 - Score	F1 - Score	F2 - Score
SVM	85.73%	83.48%	82.22%
MLP	76.83%	75.87%	75.28%
RF	85.73%	83.48%	82.22%
KNN	83.87%	81.14%	79.77%
SG	88.38%	85.33%	83.73%

3. Results

A. Precision and Recall

Basic inference of the models begins with the most fundamental of the metrics: precision and recall. The latter is taken into account along with macro and weighted averages.

Algorithm	Macro Precision	Weighted Precision	Macro Recall	Weighted Recall
SVM	88%	87%	82%	85%
MLP	78%	78%	75%	78%
RF	88%	87%	82%	85%
KNN	87%	85%	79%	84%
SG	91%	89%	83%	87%

B. F Measures

The entire model's harmonic average can be calculated by the use of precision and recall through F Measures

Algorithm	Accuracy F- Score	Macro F- Score	Weighted F- Score
SVM	85%	83%	85%
MLP	78%	76%	78%
RF	85%	83%	85%
KNN	84%	81%	83%
SG	87%	85%	87%

Fβ - scores with β values of 0.5, 1, and 2 are:

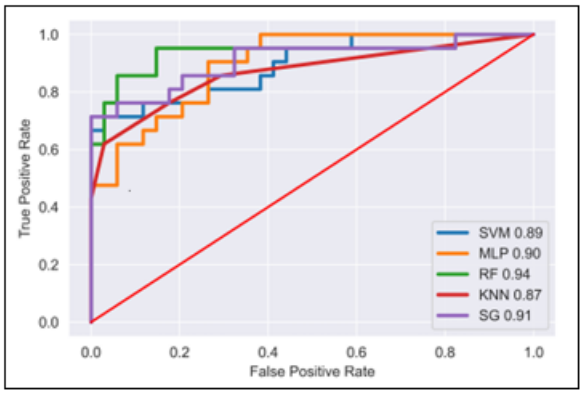
Algorithm	Hamming Loss
SVM	14.54%
MLP	21.81%
RF	14.54%
KNN	16.36%
SG	12.72%

C. Hamming Loss

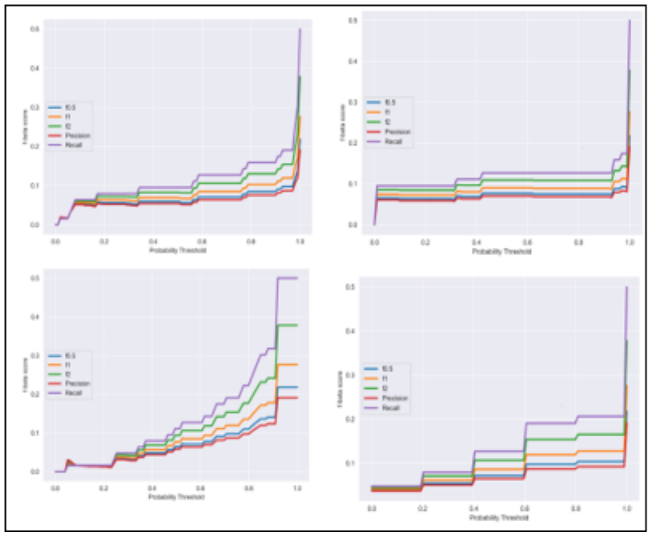
Algorithm	Jaccard Index
SVM	74.60%
MLP	64.17%
RF	74.60%
KNN	71.87%
SG	77.41%

D. Jaccard Index

Having a higher threshold of generalization criterion helps understand better opportunity within the scope of unseen data which affects performance quite significantly. Stacked Generalization seems to outrun other models in terms of F - measure which balances precision and recall while Random Forest models achieve a higher ROC - AUC score, showing an evident separation of classes within the model. Unlike broader application of algorithms where their performance will show lack of evaluation score due to enhanced generalization. In contrast, betterment of evaluation scores highlighted improved generalization rather without application of algorithm. Different metrics highlight various domains of model performance as illustrated in figure 3 which demonstrates the importance of having specific evaluation metrics relevant to the core focus of the analysis.



As various evaluation techniques will be employed in the analysis, studying F - scores in relation to their precision and recall values could be insightful. These relationships may be better demonstrated through an integrated evaluation approach, such as threshold - dependent performance curves, depicting subtle nuances of trade - off interdependencies between classification metrics within the later sections of this study. Lend me your eyes, ears, and brain for a moment as I show you what was done in this study.



4. Conclusion

Focused on a rather elegant yet undermined point in the Machine Learning field that does not receive much attention, this paper set out to address the issue. The algorithms impose strain when learning independently about the representations, especially when trying to learn from exceedingly high dimensional data with abundant categorical variables. The categorical variables, when pre - processed to numeric values, either create more features or enhance the difficulty of learning for the algorithm. Increases in complexity lead to becoming less generalizable and inducing more errors. To tackle this problem, we proposed Stacked Generalization Ensemble learning in this paper which aids in developing vague models for machine learning without transitioning to deep learning. This approach is best used when the number of

features is high and records are low. To demonstrate this, we applied the Stacked Generalization to PCOS classification, drawn comparisons with other reputedly efficient algorithms, and substantiated the claim that some metrics do not always provide genuine insights into accuracy, which is why diversification of metrics is crucial to obtain optimal results.

References

[1] Roshan Kumari and Saurabh Kr. Srivastava. Machine learning: A review on binary classification. International Journal of Computer Applications, 2017.
[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 2015.
[3] Chigozie Nwankpa et al. Activation functions: Comparison of trends in practice and research for deep learning, 2018.
[4] Abien Fred Agarap. Deep learning using rectified linear units (ReLU), 2019.
[5] Zhou, Y., et al. Self - Discover: Large Language Models Self - Compose Reasoning Structures, 2024. arXiv: 2402.03620
[6] Liu, Z., et al. KAN: Kolmogorov–Arnold Networks, 2024. arXiv: 2404.19756
[7] Tworkowski, S., et al. Buffer of Thoughts: Thought - Augmented Reasoning with Large Language Models, 2024. arXiv: 2406.04271
[8] S. Kim, H. Chu, J. Park, and J. Lee. Stacked generalization as a computational method for the genomic selection. *Frontiers in Genetics*, 2024.
[9] Z. Liu et al. A novel stacking ensemble learner for predicting residual strength of pipelines. *Nature Scientific Reports*, 2024.
[10] S. Antonik and B. Bąba. Stacked Generalization - Investigating the impact on predictive performance. *CEUR Workshop Proceedings*, 2024.
[11] Chakraborty, B. et al. (2024). A Prediction Model for Detecting PCOS Using Stacked Ensemble Learning. IEEE Xplore.
[12] Sayma, A. et al. (2024). Ultrasound - Based PCOS Detection Using Stacked Ensemble Models. IJISAE; 12 (4).