International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

An Interpretation of Stacking and Blending Approach in Machine Learning

Divya Khyani[1], Soumya Jakkula[2], Sanjana Gowda N C[3], Anusha K J[4], Swetha K R[5]

[1-4]UG Student, Dept. Of Computer Science and Engineering, BGS Institute of Technology,Karnataka, India.
[5]Assistant Professor, Dept. Of Computer Science and Engineering, BGS Institute of Technology, Karnataka, India.

Abstract - *This research paper aims to provide a general perspective on two important Machine learning approaches: Stacking, and Blending. It focuses on building up a base that helps in attaining a general idea over the technology. It explains the concept of Stacking for Classification, Stacking for Regression, Stacked Generalization and, Stacking Scikit Learn Api. In addition to that, it also gives a brief overview about concepts such as Stacking and Blending. It helps in understanding their working and, the real-life application that come under these learning techniques. At last, this research provides the comparison of stacking and blending, attempting to find which one is the best.*

Key Words: Stacking, Blending, Scikit-Learn, Classification, Regression.

1. INTRODUCTION

The collective custom nowadays is to check the reviews of items before buying them. Reviewing helps in allowing the customer's experience with a certain product be available to everyone. And while checking reviews, you often look for the items with a large number of reviews so you would know for sure about its efficiency. After looking through multiple reviews from people you decide whether to purchase the item or not.

Ensemble models in machine learning function on a similar idea. They conjoin the decisions from various models to improve the overall operation of the model. This approach provides a better predictive performance when compared to a single model and helps in accuracy. This is the reason why ensemble methods were given high priority in many significant machine learning competitions, such as the Netflix Competition, KDD 2009, and Kaggle.

Ensemble models can help confront some extremely complex machine learning problems such as overfitting and underfitting. Bagging, Boosting, Stacking, and Blending are some of the popularly known ensemble learning methods.

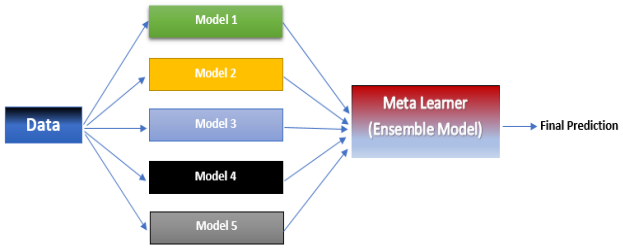

Fig -1: Ensemble Learning Technique

2. STACKING

Stacking is mostly used to combine numerous classifications or regression models into a single ensemble. It's also a different way of thinking. It is mostly used to locate a space of possible models for a given problem. Ensemble models can be created in a variety of techniques, the most well-known of which are bagging and boosting. It is primarily used to reduce variance and avoid concerns with overfitting.In supervised learning, boost is also used to lower bias and variance. The design is such that you can choose a learning model that can learn some but not all of the problem. As a result, you can create a variety of learners and utilise them to generate an intermediate prediction, one for each learnt model.

2.1 Stacked Generalization

Stacking generalisation, also known as stacking, is mostly utilised in ensemble models, where we can create a new model that will be trained to combine predictions from two or more previously trained models. Existing models or sub-models' predictions are merged using the new. Blending is another term for stacking.When the forecasts that are merged are expert, which is expert in many ways, stacking works best. We can use a basic linear approach like simple voting to merge the predictions for sub-models, or we can use a leaden sum using linear regression or logistic regression.

2.2 Stacked Scikit-Learn AP

Stacking can be done from scratch, although this will be a challenge for beginners. This Scikit-learn python machine learning library provides a framework for machine learning implementation. The StackingRegressor and StacxkingClassifier classes will be provided for stacking.

Both models will be used in the same way and will be based on the same assumptions.We must first define the list of estimators and the final estimators before we can use these models. Each model in the list might be the pipeline, along with any data preparation that the model requires before it can be idealed on the training dataset.

2.3 Stacking for Classification

Stacking can be used for the classification process as well. To fulfill the same, we use the function "make_classification()" to create a binary classification problem with some unique features and user inputs. Running these inputs would help in creating the data-set and help in understanding the shape of the components. Now, we can move ahead with the classification by using different machine learning models on the created data set. To name a few algorithms that can be used in the said mechanism- Logistic Regression,k-Nearest Neighbors, Decision Tree, Support Vector Machine, Naive Bayes, etc.

Every algorithm is used along with its default model hyper-parameters for the evaluation individually. The function that is used to create the model desired by the user is get_models(). The process used to evaluate each of these models is called repeated k-fold cross-validation and the function used is evaluate_model(). After the successful evaluation, we can also find the average performance of each classifier algorithm and also create a visualization of the same via a box and whisker plot. This will help us understand the distribution of accuracy scores for each algorithm.

2.4 Stacking for Regression

Regression is a very essential process in machine learning, which randomly takes a group of variables to predict the value Y, and attempts to build a mathematical equation between them. We can use the function "make_regression()" to create a synthetic regression problem with user inputs and a couple of features to start the stacking process. Similar to the approach mentioned in the "Stacking with Classification" section, the inputs can be executed to create the data-set and understand the shape of the components. After the creation of the dataset, the next step includes, evaluating a suite of different machine learning models. To name a few Algorithms that can be used for the evaluation of regression-k-Nearest Neighbors, Decision Tree, Support Vector Regression.

Further, the model is created with the help of the hyper-parameters given to the above-mentioned algorithms. The model performance will be predicted using the mean absolute error (MAE). The scikit-learn library is used for the sign inversion on this error to make it more accurate by turning it from -infinity to 0 to receive a better score in the end.

2.5 Applications of Stacking

Lately, because of the implementation of ensemble methods, the computational time has reduced significantly. Due to the same reason, the number of applications has increased tremendously in a couple of years that have gone by. Some of the applications are:

2.5.1 Face Recognition-Face recognition has become of the most common areas under pattern recognition for identification and verification of an image based on it saved history. It plays a great role currently in the expanses regarding mobile devices security.
2.5.2 Remote Sensing-Remote sensing obtains data from objects and occurrences without actually having to make any physical contact.
a)Land Cover Mapping- It happens to be an application of earth observatory satellite sensors using the data that is acquired from remote sensing as well as geospatial data. This is done to identify the objects and materials in the required (target) zones.
b)Change Detection- It deals with the problem related to analysis of images, these images mostly are of the ones that have had changes in the cover or surface over time. One of the main uses for change detection is to be able to assess and monitor disasters for the safety of the civilians.
2.5.3 Medical Field-The process of stacking has been successfully implemented in multiple sectors of medicine such as diagnosis for neuro cognitive disorders (neuroscience) such as Alzheimer's or myotonic dystrophy which can be attained form MRI datasets.
2.5.4 Fraud Detection-It mainly deals with discovery of money laundering, bank and credit card fraud, which also comprises of telecommunication fraud. They have enormous domains in in research and application of machine learning. Hence stacking helps in creating a model which improves the robustness of the behavior of the model and detects fraud in the banking systems.

3. BLENDING

Blending is an ensemble machine learning algorithm. It is another name for stacked generalization or stacking ensemble which stands out in terms of its fitting the meta-model approach that occurs on out-of-fold results founded by the base model, it is said to fit on predictions made on a holdout data configuration.

The blending approach is important in machine learning because it describes stacking models that combined multiple predictive models by competitors in the $1M Netflix machine learning battle. This approach is also taken under consideration when entering for stacking in competitive machine learning circles, such as the Kaggle community.

3.1 How to develop a Blending Ensemble?

At the time of writing, the scikit-learn library does not support blending constitutionally. To overcome this problem, we can use scikit-learn models for our implementation. To get started, the first step would be to create some base models. For choosing the base models we need to first understand what problem needs to be tackled regression or classification problem. After the selection of the base model, we need the list of models. A function can be defined for this purpose, get_models() since it returns a list of models. The list we can access now contains models defined as a tuple with a name and the classifier or regression object after configuration.

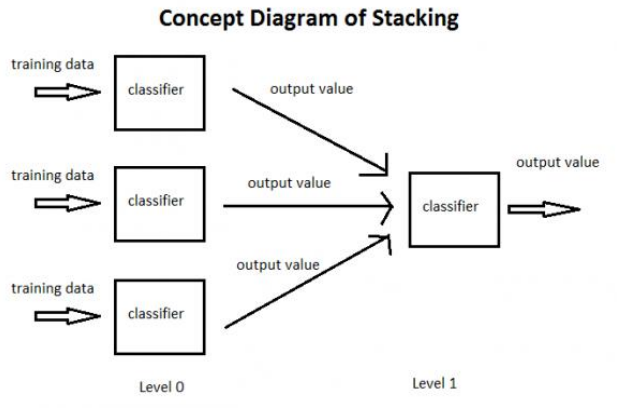

Fig -2: Concept Diagram of Stacking

For example: For a classification problem, we might use models such as logistic regression, kNN, decision tree, SVM, and Naive Bayes model. The next step will be to fit the blending model. To get started, we can jot down the list of models and fit each one by one on the training dataset. Also in this looping-like process, we can use the model that is fit to make a prediction on the validation dataset and store the predictions for future reference. The next step is to train our meta-model. We can use any machine learning model we like, such as logistic regression for the classification process. Further, we can put all of them together into a function named fit_ensemble() that helps in training the blending model using a dataset that is being trained and a dataset used in holdout validation.

After the training, we can use the blending ensemble to make predictions on new data entered by the user. This is a two-step process. The first step involves using each base model to make a prediction. These predictions are then grouped and used as input to the blending model to make the final output (giving us the prediction rate).

We can use the same looping structure as we did when we trained the model. This implies that we can collect the predictions from each base model into a training dataset, stack the predictions together, and call predict() on the blender model with this dataset at the meta-level. Finally, now we have all of the elements required to implement a blending ensemble for classification or regression predictive modeling problems.

3.2 Applications of Blending

As blending is the same as stacked generalization which uses machine learning model to learn how best to combine multiple predictions to generate a model accordingly. Some of the applications are namely:

3.2.1Financial Decision Making-In financial decision-making, the accuracy towards being able to predict failure in a business is very crucial. Therefore, blending is proposed to predict financial distress and financial crises for running a successful business. Another key feature is to be able to find out stock market manipulation.

3.2.2Emotion Recognition-Most industry players like Google, IBM and Microsoft reveal that their essential technology of speech recognition is grounded on the approach that is speech-based emotion recognition which can be achieved with blended ensemble learning .

3.2.3Computer Security-

a)Malware Detection- Computer viruses, worms, trojans etc. are classifications of malware codes with respect to machine learning, document categorization problem arises from this and hence blended ensemble method can provide proper efficiency in the area.

b)Intrusion Detection- It monitors computer networks or computer systems to identify any intruder codes like an anomaly of an event. Ensemble learning thus helps the monitoring systems to reduce their total errors successfully.

c)Distributed Denial of Service- It is one of the most threatening cyber-attacks that could take place on an internet provider. The process of blending combines the output of a single classifier and reduces the total error of detecting and discriminating of such attacks.

4. STACKING V/S BLENDING

Stacking and Blending are extremely dominant ensemble methods. They can effectively boost the model's performance and many cases can be a deciding factor to win competitions. Stacking uses out of fold predictions for the training set of the meta model, while blending uses a validation set to train the meta model (next layer). Stacking happens to be a k-fold cross validation, where the training data of the meta model is correspondent to the model-based training data. Blending is a holdout method that directly divides the training data into two parts, out of which about 10% is used for training of the metamodel. Stacking is capable of generating meta features for each sample, when the test generates meta features k weighted average is applied. Blending carries out "one-holdout set", that is, a small percentage of the training data to make forecasts which will be "stacked" to form the training data of the meta-model.

5. CONCLUSION

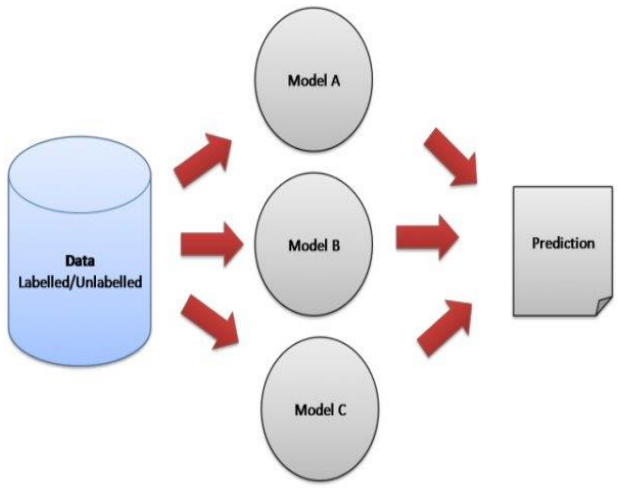

Fig -3: Overview of Ensemble Learning

Ensemble methods are methods whose purpose is to improve the accuracy of results in models by putting together multiple models instead of using a single model. The combined models increase the precision of the results significantly. This has boosted the acceptance of ensemble methods in machine learning.

Stacking was initiated by Wolpert in the paper Stacked Generalization in 1992. It is a process that uses k-fold for training base models which then make estimates on the remaining fold. These supposed out of fold predictions are further used to train another model—the meta model—which can use the data produced by these base models to provide final predictions.

Blending is the term introduced by the Netflix competition victors. It is very identical to stacking with the only variance being the use of k-fold instead of out-of-fold as predictions to create a small holdout data-set which will then be used in the training of the meta-model.

REFERENCES

[1] P.Smyth and D.Wolpert. Linearly combining density estimators via stacking. Machine Learning, 36(1-2):59-83, 1999.

[2] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[3] Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539-550.

[4] https://en.wikipedia.org/wiki/Ensemble_learning

[5] Charu C. Aggarwal, Data Classification: Algorithms and Applications, Chapman & Hall/CRC, 2014.