Stacking Supervised and Unsupervised Learning Models for Better Performance

Pulkit Kalia[1]

[1]*Engineer/Data and Decision Analyst, Mahindra Comviva/Hashtag Foods*

Abstract – *Stacking different supervised learning models have been proved to yield better accuracy and performance and with this paper, we extend the same by stacking different types of learning models, i.e. supervised and unsupervised learning models. Learning models and model stacking are increasing used today in classification, regression problems and also in image and speech recognition. The biggest advantage of Stacking is its high accuracy but may become obsolete if different models generate similar output or data is not diverse enough. One of the methods to counter this situation is presented in this paper, i.e. by stacking unsupervised learning models with supervised learning models. By stacking different types of learning models, the overall accuracy, recall and precision is increased which gives an edge to businesses working on very big or real time data.*

Key Words: **Predictive Analytics, Machine Learning, Model Stacking, Variable Reduction, Supervised learning, Unsupervised learning**

1. INTRODUCTION

Model Stacking (part of Ensemble Model) refers to using output from different Machine Learning Algorithms and using them together which results in better accuracy, recall and precision. There are mainly two types of machine learning models used in predictions and classifications:

1. Supervised learning models- Theses models have a specific target variable to predict and a set of predictors to work on (like linear models, random forest, svm, decision trees, gbm, xgboost etc.).

2. Unsupervised learning models- These models have no target variable to work upon. They check for anomaly or make groups (clusters).

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. Because of output from different models being stacked together, ensembles can be shown to have more flexibility and accuracy on the resulting model. This flexibility can, in theory, enable them to over-fit the training data more than a single model would, but in practice, some ensemble techniques tend to reduce problems related to over-fitting of the training data.

Typically, ensembles tend to give better results when there is a significant diversity among the models used for stacking. Using a variety of learning algorithms, however, has been shown to be more effective than using techniques that with similar algorithms and output.

1.1 Supervised Learning Models

The majority of practical machine learning uses supervised learning.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers; the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. Supervised learning problems can be further grouped into regression and classification problems.

1. Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

2. Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

1. Linear regression for regression problems.

2. Random forest for classification and regression problems.

3. Support vector machines for classification problems.

1.2 Unsupervised Learning Models

Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there are no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.

Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

1. k-means for clustering problems.

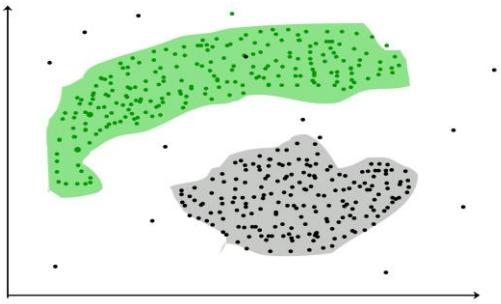
2. Apriori algorithm for association rule learning problems.

1.3 Clustering

It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



Two of the most used and common clustering techniques are:

1. Hierarchical clustering - Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

2. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

2. PREDICTIVE ANALYTICS

Predictive analytics comprises of varied statistical trends and techniques ranging from machine learning and predictive modeling to data mining to efficiently analyze the historical data and information so as to process them to create predictions about the unknown future events. As per the business aspect of predictive analytics, predictive analytics help in exploiting the patterns found in the historical business data to identify the risks and opportunities. It captures the relationships between various factors to provide the assessment of risk or a potential threat and help guide the business through important decision making steps. Predictive analytics is sometimes described in reference to predictive modeling and forecasting. Predictive analytics is confined to the following three models that outline the techniques for forecasting.

a) **Predictive Models** – Predictive models are the models that define the relationship between the various attributes or features of that unit. This model is used to assess the similarities between a groups of units providing assurance of the presence of similar attributes being exhibited by a group of similar units.

b) **Descriptive Models** – Descriptive models are the models that identify and quantify the relationships between the various attributes or features of the unit which is then used to classify them into groups. It is different from the predictive model in the ability to compare and predict on the basis of relationship between multiple behaviors of the units rather than a single behavior as is done in the predictive models.

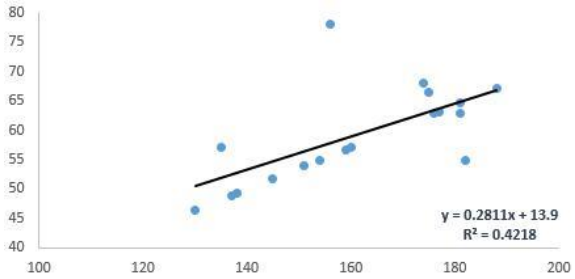
c) **Decision Models** – Decision models are the models that identify and describe the relationship among all of the varied data elements present that includes the known data set upon which the model is to be defined, the decision structure that is defined for classification and categorization of the known data set as well as the forecasted or predicted result set on the application of decision tree on the known data set so as to identify and predict the results of the decisions based on multiple attributes or features of the data set.

3. Commonly Used Machine Learning Models

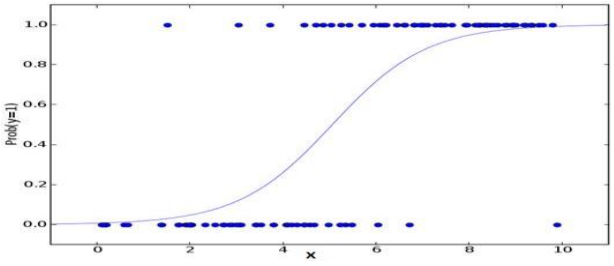
1. Regression Analytics

Regression techniques are focused on establishment of mathematical equations so as to model, represent and process the information from the available data set. Some of the regression techniques being in use are described as follows.

a. Linear Regression Model – This technique establishes a linear relationship between the dependent variable y and multiple independent variables x. It is represented through the linear equation **y=a+bx+c**



b. Logistic Regression – This technique is applied so as to find the probability regarding the success or failure of an event. This technique comes to use when the value of the dependent variable is binary.



c. **Stepwise Regression –** This technique comes into play when there is a presence of multiple independent factors or variables. The best fit is predicted through stepwise incremental addition or removal of predictor variables as required for each of the step. This technique has the aim of achieving the maximum prediction power with the use of minimal number of predictor variables.

d. **Ridge Regression –** Ridge regression technique is used where there is a multi-collinearity that is the data set has multiple independent variables with high extent or correlation. The ridge regression technique can be represented through the equation **y = a + b1x1 + b2x2 + b3x3 +...+ e**

e. **Lasso Regression –** Lasso regression is highly similar to ridge regression technique with less variance coefficients and high accuracy of the linear regression models. In this technique, variables having high correlation, only of the predictor variables is picked while all others are shirked to zero.

f. **Elastic net Regression –** This technique is a combination of Ridge Regression technique and Lasso Regression technique. It enhances the accuracy of the best fit result and provides the advantage of no limitations over the number of variables selected and has the ability to suffer and withhold double shrinkage.

2. Classification Analytics

Classification is generally used to predict where a data belong to a certain group. The predicted value is fixed (say Class A or Class B). The following models can be used for classification:

a) **Support vector machines** – SVMs are designed and defined to detect and identify the complex patterns and sequences within the data set through clustering and classification of the data. They are also referred to as the learning machines.

b) **Naïve Bayes** – Naïve Bayes is deployed for the execution of classification of data through the application of Bayes Conditional Probability [8]. It is

basically implemented and applied when the number of predictors is very high.

c) **k-nearest neighbors** – This technique involves pattern recognition techniques of statistical prediction. It consists of a training set with both positive and negative values.

d) **Random Forest**– A random forest is basically an ensemble of decision trees. Each tree classifies (often linearly) the dataset using a subset of variables. The number of trees in the forest and the number of variables in the subset are hyper-parameters and must be chosen a-priori. The number of trees is of the order of hundreds, while the subset of variables is quite small compared with the total number of variables. Random forests also provide a natural way of assessing the importance of input variables (predictors). This is achieved by removing one variable at a time and assessing whether the out-of-bag error changes or not. If it does, the variable is important for the decision.

1. Approach Taken

1. **Data Modeling –** Different types of learning algorithms are used individually and their Accuracy, Precision and Recall are calculated on the validation test set. It is advised to choose learning algorithms which shows diversity. In this way, it will be possible to harness the best of all the algorithms together later by stacking them.

2. **Model Stacking –** Outputs from different algorithms (supervised and unsupervised (clustering) learning algorithms) are collected (either as probabilities or as class names). All the outputs are transformed into a new data frame with the target variable and trained over a new variable using any other supervised learning algorithm, say knn or neural networks.

3. **Training new data frame with the resulting data frame –** The resulting data frame is trained with the resulting data frame which contains output from supervised and unsupervised learning algorithms, recall and precision increases due to the fact that different models give different outputs on the same data. Different algorithms works better with different sets of data, in this way, best outputs from all the algorithms are pooled which increases the accuracy. By including a unsupervised learning algorithm different features are grouped together to further enhance the performance.

2. Advantages

1. The resulting model is more robust than the original or individual models.

2. The accuracy, recall and precision of the stacked model are better and perform better due to the fact that different models have different outputs.

3. Stacking a unsupervised learning algorithm improves the overall accuracy more than by adding another supervised learning algorithm.

3. Disadvantages

1. Stacking can have no positive effect if the outputs from different algorithms are very close or similar.

2. Sometimes the training data has very low variability, in that case stacking can have very less or no effect.

4. Conclusion

The overall accuracy, recall and precision of the data set is increased by stacking and specially by adding output from a unsupervised leaning algorithm. A similar project has been pushed to GitHub (https://github.com/pulkitkalia1994/StackingSupervisedAndUnsupervised) which can be downloaded and extended for further research by anyone.

REFERENCES

1) Nyce, Charles (2007), Predictive Analytics White Paper(PDF), American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America, p. 1

2) Eckerson, Wayne (May 10, 2007), Extending the Value of Your Data Warehousing Investment, The Data Warehouse Institute

3) Coker, Frank (2014). Pulse: Understanding the Vital Signs of Your Business (1st ed.). Bellevue, WA: Ambient Light Publishing. pp. 30, 39, 42, more. ISBN 978-0-9893086-0-1.

4) Candemir, Sema & Antani, Sameer. (2018). A novel stacked generalisation of models for improved TB detection in chest radiographs.

5) Fletcher, Heather (March 2, 2011), "The 7 Best Uses for Predictive Analytics in Multichannel Marketing", Target Marketing

6) Barkin, Eric (May 2011), "CRM + Predictive Analytics: Why It All Adds Up", Destination CRM

7) McDonald, Michèle (September 2, 2010), "New Technology Taps 'Predictive Analytics' to Target Travel Recommendations", Travel Market Report

8) Moreira-Matias, Luís; Gama, João; Ferreira, Michel; Mendes-Moreira, João; Damas, Luis (2016-02-01). "Time-evolving O-D matrix estimation using high-speed

GPS data streams". Expert Systems with Applications. 44: 275–288. doi:10.1016/j.eswa.2015.08.048.

9) Stevenson, Erin (December 16, 2011), "Tech Beat: Can you pronounce health care predictive analytics?", Times-Standard

10) Lindert, Bryan (October 2014). "Eckerd Rapid Safety Feedback Bringing Business Intelligence to Child Welfare" (PDF). Policy & Practice. Retrieved March 3, 2016.

11) Florida Leverages Predictive Analytics to Prevent Child Fatalities -- Other States Follow". The Huffington Post. Retrieved 2016-03-25.

12) Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, Wei-Yang Lin (2009). Intrusion detection by Machine Learning : A Review

13) Siegel, Eric (2013). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (1st ed.). Wiley. ISBN 978-1-1183-5685-2.

14) Ian H. Witten and Eibe Frank. Data Mining : Practical Machine Learning Tools and Techniques.

15) New Strategies Long Overdue on Measuring Child Welfare Risk - The Chronicle of Social Change". The Chronicle of Social Change. Retrieved 2016-04-04.

16) Eckerd Rapid Safety Feedback® Highlighted in National Report of Commission to Eliminate Child Abuse and Neglect Fatalities". Eckerd Kids. Retrieved 2016-04-04.

17) A National Strategy to Eliminate Child Abuse and Neglect Fatalities" (PDF). Commission to Eliminate Child Abuse and Neglect Fatalities. (2016). Retrieved April 4, 2016.

18) Maind, S.B. & Wankar, P. (2014). Research paper on basic of Artificial Neural Network. International Journal on Recent and Innovation Trends in Computing and Communication. 2. 96-100.

19) Yuhong Yang. Aggregating Regression Procedures for Better Performance. Bernoulli, forthcoming.

20) David Wolpert. Stacked Generalization. Neural Networks, 5:241-259, 1992.

21) Leo Breiman. Stacked Regressions. Machine Learning, 24:49-64, 1996a.

22) Leo Breiman. Bagging Predictors. Machine Learning, 24:123-140, 1996b

23) Pulkit Kalia, Model Stacking: A way to reduce training time for Neural Networks, International Research Journal of Engineering and Technology (IRJET) Volume 5, Issue 9, September 2018 S.NO: 28