International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 09 | Sep 2018 www.irjet.net p-ISSN: 2395-0072

Model Stacking: A way to reduce training time for Neural Networks

Pulkit Kalia

Engineer/Data and Decision Analyst, Mahindra Comviva/Hashtag Foods

---

Abstract – *With this paper, we explore one of the ways to reduce the modeling time of a neural network. Neural Networks are increasing used today in classification, regression problems and also in image and speech recognition. The biggest advantage of Neural Networks is its high accuracy but lags behind other models in terms of processing time to train the model. One of the methods to counter this situation is presented in this paper, i.e. by model stacking. By stacking, the overall training time is reduced considerably while maintaining a similar accuracy score which gives an edge to businesses working on very big or real time data.*

Key Words: **Predictive Analytics, Machine Learning, Neural Networks, Model Stacking, Variable Reduction**

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:
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 Neural Networks

Neural networks were first developed in the 1950s to test theories about the way that interconnected neurons in the human brain store information and react to input data. As in the brain, the output of an artificial neural network depends on the strength of the connections between its virtual neurons – except in this case, the "neurons" are not actual cells, but connected modules of a computer program. When the virtual neurons are connected in several layers, this is known as deep learning.

A learning process tunes these connection strengths via trial and error, attempting to maximize the neural network's performance at solving some problem. The goal might be to match input data and make predictions about new data the network hasn't seen before (supervised learning).

Network's ultimate goal is to reduce the cost function associated with the task. Let's say stage 1 has all the input variables acting as neurons and stage 2 has few neurons randomly chosen (tunable). Every neuron from stage 1 is connected to all the neurons in stage 2 by some weight (numerical value) which denotes the strength of that link or bond. Similarly each neuron from stage 2 is either giving an output or connected to few neurons in stage 3.

1.2 Cost Function

A cost function in neural networks is a penalty awarded to the network for a wrong output. The cost function is given by:

Cost= Expected output - Predicted output

Let's suppose we have input values a_1, a_2, a_3 ….. a_n and weights w_1,w_2,w_3 …… w_n in stage 1 of the neural network then the value assigned to each neuron is given by:

$$\Theta\ (a_1 * w_1 + a_2 * w_2 + a_3 * w_3 + \ldots\ldots + a_n * w_n + \alpha)$$

Where α is a bias added and θ is the activation function (sigmoid, tanh, tan2h etc.).

The neural network then tries to minimise the cost function by backpropagation mechanism where the values are back tracked and the values of weights and biases of each neuron is altered and tested. This process is repeated again and again till a minimum cost value is reached.

1.3 Backpropagation

Backpropagation is a method used in artificial neural networks to calculate a gradient that is needed in the calculation of the weights to be used in the network. It is commonly used to train deep neural networks, a term referring to neural networks with more than one hidden layer.

Backpropagation is a special case of an older and more general technique called automatic differentiation. In the context of learning, backpropagation is commonly used by the gradient descent optimization algorithm to adjust the weight of neurons by calculating the gradient of the cost function. This technique is also sometimes called backward propagation of errors, because the error is calculated at the output and distributed back through the network layers iteratively to find the minimum cost value.

2. **Predictive Analytics**

Predictive analytics comprises of varied statistical trends and techniques ranging from machine learning and predictive modelling 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 modelling and forecasting. Predictive analytics is confined to the following three model that outlines the techniques for forecasting.

1. 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.

2. 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.

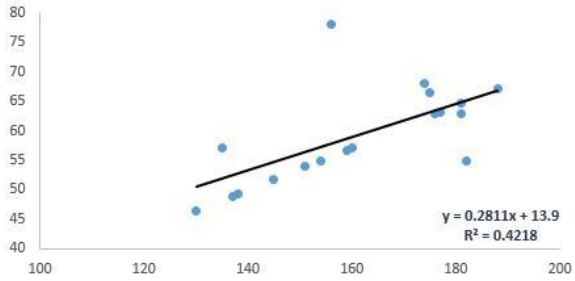
3. Decision Models – Decision models are the models that identify and describe the relationship among all of the varied data elements present that includes the known data set upon which the model is to be defined, the decision structure that is defined for classification and categorization of the known 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

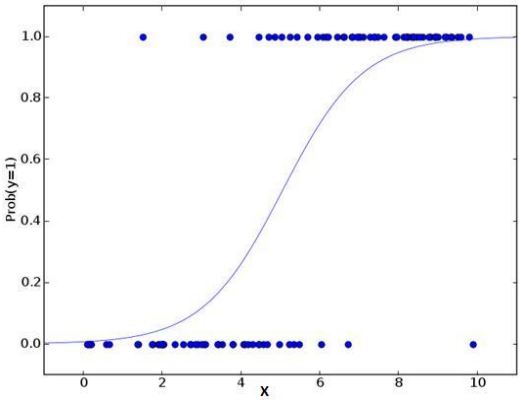
1. **Regression Analytics**

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. Polynomial Regression – In this technique, the prediction line is not a straight or linear one, but is a curve that fits the points of the data set being predicted upon.

d. 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.

e. 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

f. 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.

g. 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.

4. Approach taken to reduce training time of neural networks

a) 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.

b) Model Stacking – Outputs from different 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 neural networks. In this way initial inputs a1, a2, a3 ….. an are reduced to 5 to 6 (or close) inputs which are essentially outputs from different algorithms(random forest, ctree, knn, rpart etc.) which are way faster than training neural networks.

c) Training new data frame with neural networks – The resulting data frame is trained with neural network and the resulting accuracy, 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 and decreases the time considerable as compared to using all the input data from original data on neural network.

5. Advantages

1. The resulting model is more robust than the original or individual models.
2. The resulting model takes less time as compared to the one where all the initial inputs are used for neural networks.
3. The accuracy, recall and precision of the stacked model are better and perform better due to the fact that different models have different outputs.

6. 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.

7. Conclusion

The training time of the newly constructed data frame is considerable less than the time taken to train the data with initial data. Also, the accuracy, recall and precision is increased which serves as dual benefits. A similar project has been pushed to GitHub (https://github.com/pulkitkalia1994/SpamMessageDetection) 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