Project Review

Binary Classification of Kiva Loan Dataset using Advanced Ensemble Techniques

Joe Abley, jabley@uwo.ca

December 1, 2018

This is a review of [2], as presented in ECE 9603; the full project description and detailed results have not been made available. This review is submitted as part of the coursework for ECE 9603.

1 Project Summary

Kiva is a crowdsourced micro-lending company that makes loans to financially-excluded people, mainly in developing regions [1]. This project aims to produce a model which can predict whether the loan made to a borrower is either above or below average, based on particular characteristics of the borrower.

A substantial dataset obtained from Kiva was reduced by constraining a set of numeric features to those that were assumed to be relevant and considering only observations from loans made in a single region (India). The selected features were subsequently manipulated in order to eliminate some imbalanced classes and normalised in the range $[0, 1]$.

An ensemble learning model was chosen in order to reduce the computation time of training the model and to improve the accuracy of its predictions. The predictions of each of a two level-1 models were used as input variables for a single layer-2 model which was used to produce a set of final predictions. The two layer-1 algorithms chosen were Decision Trees and Random Forest; the layer-2 model was a multilayer perceptron with two hidden layers each of five neurons. Parameter tuning was carried out with a grid search with 10-fold cross-validation.

The results of the ensemble model were assessed using a confusion matrix, 10-fold cross-validation and F-measure and compared with the same tests applied to the component layer-1 models in order to measure the effectiveness of the stacking model. The results presented demonstrated a 3.5% increase in accuracy of the stacking model over the (presumably best-performing) layer-1 model.

2 Constructive Feedback

The problem statement would benefit from greater clarity; without prior knowledge of micro-lending practices in general or Kiva in particular, some of the commentary was a little opaque. For example, bullet-repayment was not an especially familiar concept; prior knowledge of Kiva's approach of assessing borrowers during a diligence process before they are eligible for crowdfunding would have made the problem statement easier to understand.

The way that the presentation was split between the members of the team was effective, and all members of the team spoke with authority and clarity. The content would have been easier to digest by an audience that had no prior knowledge of the project if the pace at which the slides and transitions were navigated was reduced; the presenters' enthusiasm for the project often caused them to skip ahead to material they perhaps found more interesting without always giving the audience time to absorb the more foundational details.

Some additional detail about how the team arrived at the final set of features to use as input variables for the ensemble model would have been interesting to see.

The choice of Decision Trees and Random Forest as layer-1 algorithms seems worthy of further discussion, especially since Random Forest is itself an ensemble learning method based on a multitude of decision trees. If other layer-1 algorithms were evaluated quantitatively, the evaluation process and the results would have been interesting to include in the presentation. Similarly, the process of choosing a network topology used in the layer-2 model might have been useful to include.

The layer-2 neural network itself might be a useful tool in assessing different algorithms that could have been chosen for layer-1 models. Additional layer-1 models making use of particular algorithms could be included in the aggregate ensemble model and the input layer of the neural network extended accordingly; after training, comparison of the weights applied to particular input neurons could provide a measure of the relative effectiveness of individual layer-1 algorithms.

The team's two principal goals, as described, were to increase accuracy and to reduce computation time. A modest accuracy improvement was described based on a number of accuracy metrics; it would have been useful to include the measured improvements in computation time, since significant reduction in computation would provide an additional useful illustration of why an ensemble model is a good choice to approach this problem.

References

[1] Kiva web page. https://www.kiva.org/. Accessed: 2018-12-01.

[2] K. Gupta, M. Pratapa, P. S. Vaddi, and R. Kalra. Binary classification of kiva loan dataset using advanced ensemble techniques. ECE-9603 Presentation 2018-11-29, November 2018.