Ensemble Modeling with SPSS Modeler

Zhen Zhang[1], Lei Zhang[2], James Veillette[1], Timothy Tate[1]
[1]C Spire, 1018 Highland Colony Parkway, Ridgeland, MS 39157
[2]Mississippi State Dept. of Health, 570 E. Woodrow Wilson Dr., Jackson, MS 39216

Abstract

In the predictive analytics world, ensemble modeling strategy is often pursued to improve model accuracy and robustness. Ensemble modeling is the process of creating multiple models and incorporating them into a single scoring algorithm. The value of ensemble modeling for enhancing predictive accuracy and increasing model stability is widely recognized. However, it is an art that is not easily mastered. Through our predictive modeling practice within the telecommunication industry, we have found that in general, heterogenous ensemble modeling produces better results than homogeneous built-in ensembles. We have also found that depending on the size of the modeling target, homogeneous ensemble modeling may not always be the best choice, compared to a single model. In this paper, we present empirical ensemble strategies and suggest best practices pertaining to modeling techniques and target sizes.

Key Words: predictive modeling, algorithm, ensemble, homogeneous ensemble, heterogeneous ensemble, boosting, bagging

1. Introduction

SPSS modeler provides a variety of modeling algorithms, such as Decision Trees, Neural Net, Regressions, etc. Many of the algorithms are loaded with built-in ensemble functionalities, namely, boosting and bagging[1].

Boosting ensemble is used to improve model accuracy by decreasing bias. Boosting works by building multiple models in a sequence. After the first model is built, a second model is constructed on residuals - records that were misclassified by the first model. In the same way the third model is built on the second model's errors, and so on, until convergence occurred. Finally, cases are classified by applying the entire set of models to them, using a weighted voting procedure to combine the separate predictions into one overall scoring algorithm[1]. Boosting can significantly improve the overall model accuracy, but there are also downsides, such as longer training, decreased interpretability, overfit, etc.

Bagging ensemble, also called bootstrap aggregating, is chosen to improve the stability of the model by variance reduction[2]. Bagging works by building models using random subsets of training data, then all models are given same voting power to ensemble the scoring algorithm.

Homogenous ensemble refers to the built-in boosting or bagging ensemble, as it incorporates models produced by the same type of classifier. Heterogenous ensemble, on the other hand, refers to ensemble of models produced by different type of classifiers, such as Decision Tree model with Logistic Regression model, Decision Tree model with Neural Net, model, Regression model with Neural Net model, etc..

2. Method

2.1 First, build a single CHAID tree, then build homogeneous CHAID ensembles via the SPSS built-in boosting and bagging functions, using the same input fields as the single tree; Lastly, build Logistic regression model using the same input fields.

2.2 Repeat 2.1 for N times, using different datasets. We've used real operational data from different business seasons. Unlike simulated datasets, un-edited operational data add an additional layer of testing and validation for the consistency of results.

2.3 Evaluate the accuracies of single models and boosting and bagging ensembles.

2.5 Ensemble the single CHAID tree model with Logistic regression model for each set of models. Evaluate the accuracies of the heterogeneous ensembles.

2.6 Apply compare means tests

3. Results

Table 1 shows that homogeneous ensemble significantly improves the model's overall accuracy.

Table 1. Overall Accuracy: Homogeneous ensemble vs. single model

	N	Mean	p value
Single Decision Tree	50	73.6%	
Homogeneous Ensemble	50	76.2%	<0.05

Table 2 shows that Heterogenous ensemble also significantly improves the model's overall accuracy. Furthermore, the extent of overall accuracy increases caused by these two types of ensembles are comparable.

Table 2. Overall Accuracy: Heterogenous ensemble vs. single model

	N	Mean	p value
Single Decision Tree	50	73.6%	
Heterogenous Ensemble	50	76.5%	<0.05

Target accuracy reflexes the model's ability to correctly classify the outcome of interest. Table 3 show that there is no significant difference between the target accuracy of homogeneous ensemble and single model.

Table 3. Target Accuracy: Homogeneous ensemble vs. single model

	N	Mean	p value
Single Decision Tree	50	71.5%	
Homogeneous Ensemble	50	71.7%	>0.1

Table 4 shows that Heterogenous ensemble significantly improves the model's target accuracy.

Table 4. Target Accuracy: Heterogenous ensemble vs. single model

	N	Mean	p value
Single Decision Tree	50	71.5%	
Heterogenous Ensemble	50	76.2%	<0.05

4. Discussion/Conclusion

Binary classifier can be extremely unbalanced (i.e. target very small), as seen in fraud detection and churn retention cases. In such cases the model evaluation cannot rely on the overall accuracy alone, rather, target accuracy is a more relevant measure. In this study, the comparison of the effectiveness of two types of ensembles, homogeneous and heterogeneous, are based on both overall accuracy and target accuracy. Heterogeneous ensemble is demonstrated a superior strategy, in that it not only improves overall accuracy, but more importantly, significantly improves target accuracy.

Even though homogeneous ensemble improves the overall accuracy, it does not improve target accuracy, therefore it does not practically improve the usefulness of model in its ability to identify target. In cases of small target, the added process time and added complexity of homogeneous ensemble out weighs its contribution to the overall model accuracy.

In practice, logistic regression and decision tree models are good pairs for ensemble, Neural net models also pair well with decision trees.

References

1. Decision Tree Node Building Options. Retrieved May 23, 2018 from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/trees_cart_objective.htm

2. Peter Buhlmann. Bagging, Boosting and Ensemble Methods. Chapter from book Handbook of Computational Statistics: Concepts and Methods. 2012

Acknowledgements

This work is supported by C Spire, a telecommunication and technology company headquartered in Ridgeland, Mississippi.

Submitting author:

Zhen Zhang, Ph.D., Department of Marketing, C Spire.

1018 Highland Colony Parkway, Ridgeland, MS 39157, USA.

Phone (601) 540-7157

E-mail: zzhang@cspire.com