John Hancock

Advanced Data Mining and Machine Learning CAP 6778

September 11^{th}, 2014

Review of Learning with Limited Minority Class Data

Learning with Limited Minority Class Data[1] is an investigation with the goal of finding the optimal ratio of majority to minority class membership in a dataset with respect to classifier performance.

The paper is about working with data where we have small percentage of positive class. It looks at the question of how many examples from abundant class should be used.

The results will show that 2:1 or 3:1 in favor of majority yields better results.

project used Java & Weka framework

The experiments described in the paper use 11 classification techniques on 10 datasets. Some datasets have classes combined to reduce multiple classes to two

AUC is evaluation metric for classifier performance.

Part of the experimental method is to vary percentage of positive vs. negative classes and evaluate classifiers.

The presenter, Dr. Khoshgoftaar, points out that there is an application to marketing.

Results show best AUC values for %P=35 or %P=25. For 3 out of 4 cases, the winner is %P=35, but results are close for both %P=25 or %P=35.

Another section of results covers an ANOVA. The ANOVA is on 3 factors: learners, percent of positive, and number of positive, and the interactions. Response is the AUC value. P values show all factors and interactions are significant.

Main factor #P HSD shows number of positive cases implies higher AUC Main Effect %P shows no significant difference between 25% and 35% %P

For higher rarity, #P=5, optimal ratio is 3:1 but 2:1 gives only slightly lower performance. Results also show, 1:1 is not optimal.

Future work proposed in [1] : expand number and type of datasets, look into even higher levels of imbalance.

References

[1] T. M. Khoshgoftaar et. al., “Learning with Limited Minority Class Data,” Florida Atlantic University, Boca Raton, FL, USA