THREE METHODS FOR HANDLING IMBALANCED DATA FOR CLASSIFICATION WITH DECISION TREES

Annie K. Lamar, Lia Chin-Purcell, and Sarah Walling-Bell
Department of Computer Science
University of Puget Sound, Tacoma, WA 98416
kalamar@pugetsound.edu, lchinpurcell@pugetsound.edu, swallingbell@pugetsound.edu

May 13, 2019

1 Introduction

In this paper we investigate the impact of pruning and the effects of three different methods for handling imbalanced data on a classification task with decision trees. In particular, we use grammatical attributes of the titles of works of modern art to predict artists' genders. The works and artists are taken from the metadata provided by the Museum of Modern Art.

The automatic generation of metadata allows for more efficient processing of large datasets and increased accessibility of those datasets for users. Our project investigates the plausibility of the automatic generation of metadata for works of modern art. In addition, several recent studies have suggested that the writing style in formal essays, blogs, and emails may be used to predict author characteristics such as gender. We explore to what extent such prediction is possible by using an extremely small piece of text in our classification task: titles.[1]

We evaluate the performance of both post-pruned and unpruned decision trees trained with data balanced using the methods of oversampling, undersampling, and ensemble sampling. The best performing tree was a post-pruned decision tree trained with data balanced using ensemble sampling.

2 Related Works

Decision trees are used in a wide variety of contexts. Commonly, decision trees are used in medical fields to predict diseases in patients given various attributes such as their age, ancestry, and sex. Al-Dlaeen and Alashqur (2014) use a decision tree to predict Alzheimer's disease in patients given five characteristics about the patients [4].

The topic of gender classification has also been explored with a wide variety of methods. Levi and Hassner (2015) investigate the use of a neural network to perform both age and gender classification based on photographs [5]. The researchers motivated this study by discussing how facial recognition was fairly poor when guessing age and gender. These researchers were able to train a neural network that was more accurate than the current benchmark.

Our project is particularly interested in gender classification based on text. Most existing work on gender classification has been done on formal writings, which differ from blog posts in style, grammar, and noise level (errors, abbreviations, etc.). Mukherjee and Liu (2010) investigate how to improve gender classification of blog authors [6]. The authors find that the combination of two novel approaches, including a new feature-learning algorithm, produces results with significantly higher accuracy in a gender classification task than current models. Goswami et al. (2009) also perform gender classification of blog-post authors based primarily on the authors' sentence length and use of slang [7]. Otterbacher (2010) explores gender classification in a similar genre: movie reviews. Otterbacher is able to predict the gender of authors with a 73.7

[1]Regarding gender classification, the authors would like to note that we are performing a binary classification. This is because the two genders in the MOMA dataset are "male" and "female." We affirm our support, compassion, and respect for all people who identify as transgender, non-binary, gender non-conforming, or intersex. In addition, we do not believe there is a relationship between sex and writing style; rather, we hold that writing style is an aspect of social gender performance. For more information, see [1]-[3].

3 Datasets

3.1 Data Format

We use data from the Museum of Modern Art's published metadata [9]. From the artworks data file, we keep only the title of the artwork and the artist's gender. For each example, we use the Stanford Natural Language Toolkit (NLTK) to assign part-of-speech tags to each word in the title [10]. We then count the number of nouns, number of foreign words, number of prepositions, number of determiners, and number of adjectives. These counts are the attributes which compose our training data. For the nouns attribute, we bin any count greater than 20 as 20 nouns.

Dataset (female:male)	Lines
Baseline (18:82)	5845
Oversampled (0.25:0.75)	5938
Oversampled (0.50:0.50)	9133
Oversampled (0.75:0.25)	1402
Undersampled (0.25:0.75)	3960
Undersampled (0.50:0.50)	1924
Undersampled (0.75:0.25)	1402
Ensemble (x 10)	1553

3.2 Methods for Handling Imbalanced Data

Oversampling We perform random oversampling of the minority class of examples (female artists). We randomly replicate examples with female artists as the output to produce three datasets with female:male ratios of 0.25:0.75, 0.50:0.50, and 0.75:0.25. The goal of oversampling is to make the decision tree more aware of the minority class. The main disadvantage of random oversampling is it can increase the likelihood of overfitting [11], [12].

Undersampling We perform random undersampling of the majority class of examples (male artists). We randomly remove examples with male artists as the output to produce three datasets with female:male ratios of 0.25:0.75, 0.50:0.50, and 0.75:0.25. The goal of undersampling is to make the decision tree more sensitive to the minority class [12]. The main disadvantage of random undersampling is that it discards examples that are potentially useful [13].

Ensemble Sampling We perform random ensemble sampling. This involves creating k datasets (ensembles) which include the entirety of the minority class examples and 100/k of the majority class examples. We choose k = 10. We train a decision tree with each ensemble and average the results. This approach is our own, but is inspired by Galar et al.'s work on ensemble classifiers for imbalanced data [14].

4 System Description

For each of the datasets above, we train both an unpruned and post-pruned decision tree. A decision tree takes in a set of examples; each example contains several attributes and a single, corresponding output value. We decide what attribute to split on by maximizing information gain. In this case, the ideal attribute to split on is one that divides the examples perfectly and results only in leaf nodes. Information gain is based on entropy, a measure of uncertainty:

$$InformationGain(A) = B(\frac{p}{p+n}) - Remainder(A) \tag{1}$$

where p is the set of positive examples, n is the set of negative examples, and B is defined as

$$B(q) = -(qlog_2q + (1-q)log_2(1-q)) \tag{2}$$

B represents the entropy of the entire set. The remainder represents the entropy of the set after selected attribute A and is defined as

$$Remainder(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B(\frac{p_k}{p_k + n_k}) \tag{3}$$

4.1 Performance-Based Post-Pruning

Our project explores the impact of minimum error post-pruning. This is based on the work originally done by Cestnik and Bratko (1991) and refined by Patel and Upadhyay (2012) [15], [16]. We train a decision tree. Then we look at the set of all twigs in the tree (i.e. nodes whose children are all leaves). For each twig, we calculate the accuracy on a validation set of examples (1) if the tree remained the same and (2) if that node was made a leave node and assigned the plurality value of its children. We continue looking at the tree's twigs until pruning no longer improves performance on the validation set. We evaluate performance on the validation set with accuracy of classified examples.

5 Results

We evaluate our model by calculating (1) the k-fold accuracy of the model on training data, and (2) the test set accuracy of the model on test data never used previously in training. K-fold cross validation is performed by dividing the training data into k subsets, and training the tree on subsets 1, 2, ... i-1, i+1, ... k, where i is a different subset in each iteration. We then evaluate the tree by classifying the examples in subset i and calculating the accuracy. We take the average accuracy over k iterations. We provide an overview of our results below. A full results table is included in Appendix A.

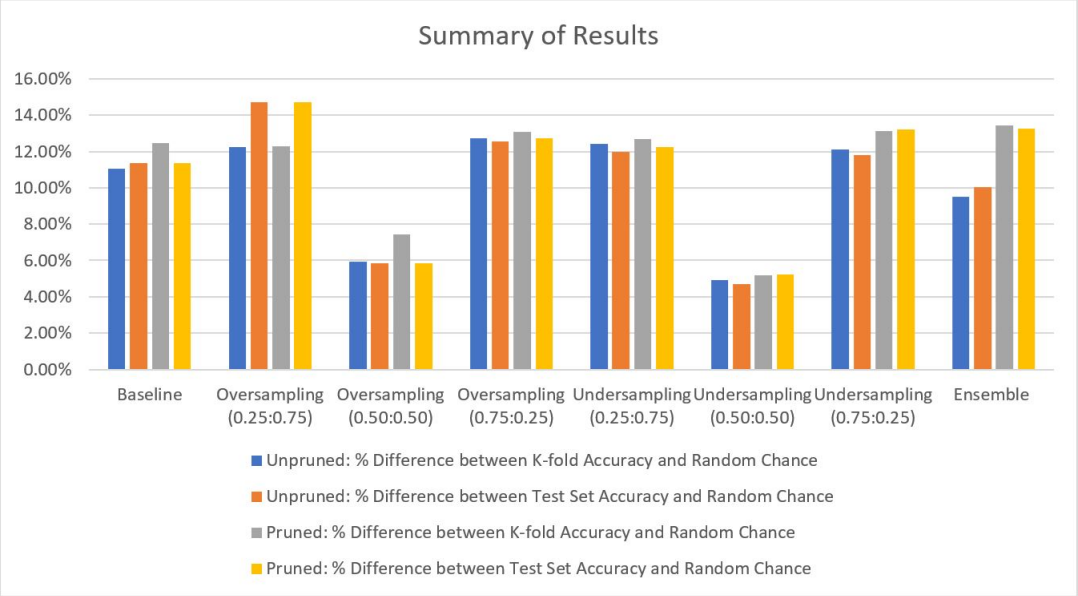

Figure 1 above shows the results for our unpruned trees. The results are presented as the difference in accuracy percentage between the attained k-fold or test set accuracy and random chance. Random chance was calculated as follows:

$$RandomGuessing = P(classisMale) * P(guessedMale) + P(classisFemale) * P(guessedFemale) \quad (4)$$

We assume that the guess is made with knowledge of the female:male ratio of the dataset for which we are calculating random chance. In general, our k-fold accuracy and test set accuracy are the same. Some interesting results are that for both oversampling and undersampling, a 0.50:0.50 ratio of females:males produced particularly poor pruned and unpruned decision trees. The k-fold and test set accuracy, respectively, for the oversampled 0.75:0.25 tree is 75.21% and 75.05%. This suggests that the tree was overfitted since guessing "female" every time would result in an accuracy of 75%. In addition, we note that the test set accuracy on oversampling 0.25:0.75 for both pruned and unpruned trees is the highest percentage attained; this accuracy is not reflected by the corresponding k-fold accuracy, however. Post-pruning had the greatest impact on the ensemble dataset, where the difference between test set accuracy and random change improved from 10.02% to 13.25%.

6 Conclusion and Future Work

We trained both unpruned and post-pruned decision trees to evaluate the effect of oversampling, undersampling, and ensemble sampling on an imbalanced dataset to classify artists' genders. Based on our results, we recommend using ensemble sampling and training along with post-pruning for the best results. We confirm the findings of other authors that oversampling results in overfitting the decision tree. We also find that it is possible to predict the gender of artists based on the title of the artwork with an accuracy slightly above random chance.

Future work on this project may include the inclusion of more attributes in training data or the exploration of other ways to handle imbalanced data. For example, Chawla et al. (2002) proposes a method to combine oversampling and undersampling [12]. In addition, future work may involve training different types of classifiers such as naive bayes classifiers and artificial neural networks.

7 Appendix A

Information			NOT PRUNED				PRUNED				PRUNING IMPROVEMENT	
Dataset	Lines	Random	K-fold	Rcdiff	Test set	Rcdiff	K-fold	Rcdiff	Test set	Rcdiff	K-fold improvement	Test set improvement
all(18:82)	5,845	70.48%	81.54%	11.06%	81.85%	11.37%	82.97%	12.49%	81.85%	11.37%	1.43%	0.00%
Oversampling ("Increased" datasets) female:male											1.43%	0.00%
0.25:0.75	5938	62.50%	74.74%	12.24%	77.23%	14.73%	74.78%	12.28%	77.23%	14.73%	0.04%	0.00%
0.50:0.50	9133	50.00%	55.92%	5.92%	55.86%	5.86%	57.42%	7.42%	55.86%	5.86%	1.50%	0.00%
0.75:0.25	18688	62.50%	75.21%	12.71%	75.05%	12.55%	75.59%	13.09%	75.21%	12.71%	0.37%	0.16%
Decreased (undersampling) female:male											0.64%	0.05%
0.25:0.75	3960	62.50%	74.92%	12.42%	74.49%	11.99%	75.17%	12.67%	74.75%	12.25%	0.25%	0.25%
0.50:0.50	1924	50.00%	54.91%	4.91%	54.69%	4.69%	55.20%	5.20%	55.21%	5.21%	0.29%	0.52%
0.75:0.25	1402	62.50%	74.60%	12.10%	74.29%	11.79%	75.63%	13.13%	75.71%	13.21%	1.03%	1.43%
Ensemble resampled datasets											0.52%	0.73%
Ensemble 1	1554	57.45%	65.32%	7.87%	73.55%	16.10%	70.22%	12.77%	73.55%	16.10%	4.89%	0.00%
Ensemble 2	1554	57.45%	68.49%	11.04%	67.74%	10.29%	71.87%	14.42%	67.74%	10.29%	3.38%	0.00%
Ensemble 3	1554	57.45%	67.63%	10.18%	65.16%	7.71%	72.52%	15.07%	67.10%	9.65%	4.89%	1.94%
Ensemble 4	1554	57.45%	70.00%	12.55%	72.90%	15.45%	70.43%	12.98%	72.90%	15.45%	0.43%	0.00%
Ensemble 5	1554	57.45%	68.13%	10.68%	73.55%	16.10%	71.01%	13.56%	75.48%	18.03%	2.88%	1.94%
Ensemble 6	1554	57.45%	64.89%	7.44%	49.03%	-8.42%	71.65%	14.20%	67.74%	10.29%	6.76%	18.71%
Ensemble 7	1554	57.45%	67.05%	9.60%	69.03%	11.58%	70.22%	12.77%	69.68%	12.23%	3.17%	0.65%
Ensemble 8	1554	57.45%	63.31%	5.86%	65.16%	7.71%	70.07%	12.62%	68.39%	10.94%	6.76%	3.23%
Ensemble 9	1553	57.49%	65.47%	7.98%	69.03%	11.54%	71.51%	14.02%	74.19%	16.70%	6.04%	5.16%
Ensemble 10	1315	70.35%	82.03%	11.68%	82.44%	12.09%	82.46%	12.11%	83.21%	12.86%	0.42%	0.76%
Ensemble Avg	1530	58.74%	68.23%	9.49%	68.76%	10.02%	72.20%	13.45%	72.00%	13.25%	3.96%	3.24%

References

[1] Judith Butler."Subversive Bodily Acts, IV Bodily Inscriptions, Performative Subversions." *Gender Trouble: Feminism and the Subversion of Identity.* New York: Routledge. 1990.

[2] Candance West and Don. H. Zimmerman. "Doing Gender." *Doing Gender, Doing Difference; Inequality, Power, and Institutional Change*, edited by Sarah Fenstermaker and Candace West. New York: Routledge, 3-25. 2002.

[3] R.F. Levant and K.M. Alto. "Gender Role Strain Paradigm." *The SAGE Encyclopedia of Psychology and Gender*, edited by Kevin L. Nadal, 718. 2017.

[4] D. Al-Dlaeen and A. Alashqur. "Using Decision Tree Classification to Assist in the Prediction of Alzheimers Disease." *2014 6th International Conference on Computer Science and Information Technology* (CSIT). 2014.

[5] Gil Levi and Tal Hassner. "Age and Gender Classification Using Convolutional Neural Networks." IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the *IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR). 2015.

[6] B. Liu and A. Mukherjee. "Improving Gender Classification of Blog Authors." *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing* (ACM), 207-217. 2010.

[7] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. "Stylometric Analysis of Bloggers' Age and Gender." *Third International AAAI Conference on Weblogs and Social Media*. 2009.

[8] Jahna Otterbacher. "Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content, and Metadata." *Proc. 19th ACM International Conference on Information and Knowledge Management* (CIKM '10). 2010.

[9] Robot, Open Data. *MoMA Collection - Automatic Monthly Update* [Data set]. Zenodo. http://doi.org/10.5281/zenodo.2655998. 2019.

[10] Steven Bird, Edward Loper and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc. 2009.

[11] M. Kubat and S. Matwin. "Addressing the curse of imbalanced training sets: One sided selection." *Proceedings of the Fourteenth International Conference on Machine Learning*, 179-186. 1997.

[12] N.V. Chawla, L.O. Hall, K.W. Bowyer, and W.P. Kegelmeyer. "SMOTE: Synthetic Minority Oversampling Technique." *Journal of Artificial Intelligence Research*, 16:321- 357. 2002.

[13] S.B. Kotsiantis, P.E. Pintelas, and D. Kanellopoulus. "Handling imbalanced datasets: A review." *GESTS International Transactions on Computer Science and Engineering* 30. 2006.

[14] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C; Applications and Reviews*. 2011.

[15] Bojan Cestnik and Ivan Bratko. "On Estimating Probabilities in Tree Pruning." *Proc. European Working Session on Machine Learning* (EWSL), 138-150. 1991.

[16] Nikita Patel and Saurabh Upadhyay. "Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA. *International Journal of Computer Application* 60.12, 20-25. 2012.