IMPACT OF NON-PROPORTIONAL TRAINING SAMPLING OF IMBALANCED CLASSES ON LAND COVER CLASSIFICATION ACCURACY WITH SEE5 DECISION TREE

Zhengwei Yang, Claire G. Boryan

National Agricultural Statistics Service (NASS)
United States Department of Agriculture (USDA)

ABSTRACT

The accuracy of a supervised classification is highly dependent upon the training samples. This paper is concerned with the impact of non-proportional training data sampling of imbalanced classes on land cover classification accuracy, using a See5 decision tree classifier. The purpose of this paper is 1) to examine experimentally how the training sampling ratio affects classification accuracy in the imbalanced class scenario; and 2) to determine the best training data sampling ratio for optimal classification performance using a See5 decision tree classifier. To better measure classification accuracy, we propose a balanced accuracy measure of a targeted class, which incorporates both False Positive and False Negative errors to truthfully reflect the accuracy of a targeted class. The study result indicates that balancing the training sample between classes does not necessarily improve the classification accuracy. Instead, selecting a training sample ratio which equals the actual ratio of the coverages of the imbalanced classes will yield the best classification performance.

Index Terms— See5/C5, class balancing, targeted class accuracy measure, best training sampling ratio

1. INTRODUCTION

Supervised classification methods, such as machine learning methods See5, Random Forest, Neural Networks and Support Vector Machine are widely used for land cover classification [1]. However, classification accuracy varies as a function of a range of training set properties [2]. Therefore, properly designing a sampling scheme and selecting a training data set are critical to the resulting classification accuracy. Properly assessing the classification performance with respect to the sampled training data is also necessary for improving classification accuracy.

It was found that in machine learning classification, class imbalance (i.e., one class population is much smaller than others) may hinder the performance of the standard machine learning classifiers [3]. The accuracy performance issue caused by class imbalance is mainly because most standard machine learning algorithms are accuracy driven. This means that many classification algorithms try to minimize the overall error (i.e. to maximize the overall classification accuracy). However, in a class imbalance dataset, overall classification accuracy tells very little about the minority class [7]. The minority class features are often treated as noise and are ignored. This leads standard machine learning classifiers to bias towards majority classes which usually have a larger number of training instances. Consequently, the minority class tends to have a lower classification accuracy as compared to the majority classes.

Research on classification of the imbalanced classes has been growing rapidly, and a number of methods have been developed recently [4][5][6][7]. In general, there are two categories of methods to handle class imbalance classification: 1) the data-level approach and 2) the algorithm-level approach. The data level approach adjusts the class imbalance ratio with the objective to achieve a balance distribution between classes whereas in the algorithm-level approach, the conventional classification algorithms are fine-tuned to improve the learning task especially related to the smaller class [7]. The data level approach for balancing classes is to either increase the minority class sample size or decrease the majority class sample size. The goal is to achieve the same number of training instances for both classes.

In this study, we use a real world scenario, the Florida citrus land cover classification, as a case study. Florida citrus crop production, a $3.34 billion industry, accounts for 49% of total U.S. citrus production. It is important to timely monitor and assess Florida citrus production for informed business decision making and policy formulation. Machine learning based classification analysis using remote sensing data provides an effective, efficient and low cost approach for monitoring such large scale crop production as the Florida citrus crop [1]. In this study, all other crop classes and non-crop classes are merged into one “other” class and the classification becomes a binary classification problem.

It is obvious that the citrus class and “other” class are not in balance since the citrus crop land cover is only about 3.2% of total land cover within the study area as calculated from historical record. Overall classification accuracy will be dominated by the “other” class (about 97% land cover) accuracy. The citrus classification accuracy will be greatly affected by the training sample balance. Therefore, this is a perfect case to study the impact of balancing classes on the classification accuracy of the minority class.

The purpose of this study is 1) to examine experimentally how the training sample ratio significantly affects classification accuracy in the imbalanced class scenario; and 2) to determine the best training data sampling ratio for optimal classification performance using a See5 decision tree classifier. This study will experimentally verify whether balancing classes improves classification accuracy of the minority class using a See5 decision tree classifier.

The rest of the paper is organized as follows: Section 2 describes study area and test data. Section 3 introduces the study methodology. Section 4 presents experiment results and discussion. Finally, Section 5 presents our conclusions.

2. STUDY AREA AND DATA

2.1 Study Area

This study uses a See5 decision tree classifier to identify the 2017 citrus crop in Florida, U.S. from Sentinel-1A SAR and Landsat 8 data. The study area for this research covers the major Florida citrus growing area, a region within Florida, which is approximately 6357 km^2, as shown in Fig. 1(red polygon). This region is selected due to its major citrus production and persistent clouds to evaluate the SAR data. Though Florida is one of the major citrus growing areas, the citrus land coverage is just about 3.2% of the study area. There are many other varieties of crops in the area. This makes it a perfect scenario for studying the impact of decision tree classification training data sampling ratio on classification accuracy.

2.2 Satellite Imagery

The geospatial data used in this assessment include: Sentinel-1A and Landsat 8 imagery, a Florida citrus mask, citrus field polygon data, and the 2011 National Land Cover Data Set (NLCD).

The ESA Sentinel-1 constellation has two polar-orbiting C-band SAR satellites Sentinel-1A and Sentinel-1B. Sentinel-1A level-1 data products, acquired on March 26, April 7, May 3 and July 24, 2017, are used in this study. Sentinel-1A's interferometric mode has 250 km wide swath. The Level-1 products are multi-look ground range detected (GRD) and have 5x20 meter spatial resolution and dual polarization (VV and HH).

Landsat 8 30m OLI Level-1 imagery used for this assessment were acquired on April 7 and May 9, 2017. The same date and path scenes were mosaicked. The bands used for this assessment include: bands 3 (visible green), 4 (visible red), 5 (near infrared), 6 (short wave infrared – 1), 9 (Cirrus) and 10 (Thermal Infrared – TIRS-1).

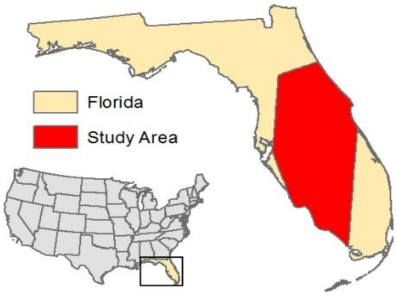

Fig. 1. Study area within Florida, U.S. (red polygon)

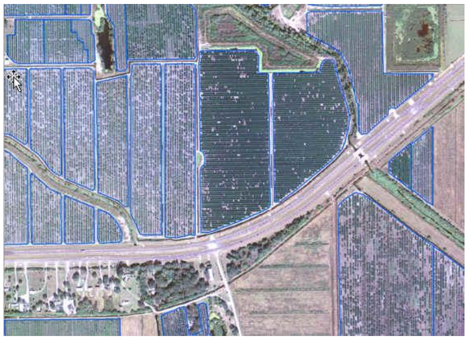

Fig. 2. Zoom of citrus polygon data used for training and validation. All polygons are manually delineated and attributed, based on annual field inspections, in the USDA NASS Florida Field Office. Blue boundaries surround the citrus groves overlaying the aerial imagery.

2.3 Citrus Mask

A 30m resolution Florida citrus mask, derived from 2013 – 2016 NASS historic Cropland Data Layers [1], is used as an ancillary layer in the classification. The citrus mask defines a pixel as citrus if the pixel is identified as citrus in all four years.

2.4 Ground Reference Training and Validation data

Citrus polygon data provided by the USDA NASS Florida Field Office are used for training and validation of the citrus class in this study. Fig. 2 shows a zoom of a portion of the citrus polygon data, outlined with blue boundaries, which are overlaid on aerial photography. All citrus groves are manually delineated and updated yearly based on field inspections which take place from October through June. There are 26,536 citrus groves recorded in the 2017 Citrus Geographic Information System (GIS) Data Layer. The citrus polygon data are the most accurate, current and comprehensive delineation of the Florida citrus crop available [8].

The 2011 U.S. Geological Survey, National Land Cover Data Set (NLCD) are used for training of all non-citrus categories of land cover [9]. The NLCD data have a 16-class land cover category scheme. It covers CONUS at a spatial resolution of 30 m.

3. METHODLOGY

3.1 Sentinel-1A and Landsat 8 preprocessing

The downloaded Sentinel-1A images were first preprocessed with calibration to sigma naught, Range Doppler terrain correction and de-speckling (median 5x5 speckle filter) using the ESA open source Sentinel-1 toolbox. The preprocessed same date images were mosaicked, reprojected to Albers Conical Equal Area projection, resampled to 30 m and set to the map extent of the study area using Hexagon's ERDAS Imagine 2016 software.

Landsat 8 OLI Level 1 scenes were reprojected to Albers Conical Equal Area Projection, mosaicked (same date and path) and set to the extent of the study area. The six bands selected in this assessment are the bands identified as most useful in classifying crops by the NASS CDL program.

3.2 See5 Classification Test and Training Sample Selection

In this study, See5 decision tree method (version 2.08), with the boosting option, is used to produce the citrus classifications. Both multi-date optical and SAR imagery are stacked for classification. A citrus mask is used as an ancillary layer for all classifications. The algorithm of the decision tree classifier is untouched. The input imagery, decision tree classifier parameter settings are unchanged for all tests.

To test the impact of class balancing on citrus classification accuracy, we sample a series of training data sets with different class sample ratios. The training data sets are randomly sampled from the Citrus GIS Data Layer for citrus training and NLCD for non-citrus training respectively. The number of sampled training data is determined by testing the sample ratio of the citrus class to "other" class with a combined total sample number of 1,000,000 points. The experimental training sampling ratios to be tested are 2%, 3%, 5%, 10%, 20%, 30%, 40% and 50%, of which the 50% ratio is the balanced class sampling ratio. For example, for the 3% ratio, there are 30,000 citrus pixels and 970,000 non-citrus pixels sampled from the Citrus GIS Data Layer and NLCD respectively. For citrus classification validation the complete Citrus GIS Data Layer including all polygons are used in this study.

3.3 Accuracy Assessment

Classification accuracy is usually measured using a confusion matrix which contains information about the actual and the classified class as shown in the following table:

TABLE 1. Binary classification confusion matrix

	Classified Positive	Classified Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Common performance measures (indices) can be derived from the confusion matrix as follows [10]:

$$Overall\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate\ (TNR) = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

True Positive Rate sometimes is also called user's accuracy (corresponds to error of commission) while Precision is also called producer's accuracy (corresponds to error of omission). Each type of accuracy measure yields different information.

Either the True Positive Rate or Precision, or the True Negative Rate measures only one performance aspect of the classifier. They do not provide a comprehensive measure of the performance. For example, a classifier that has higher precision may have a very high false negative rate. Therefore, if we focus on only one measure, we may get a biased measure of accuracy.

On the other hand, total accuracy is a summary measure, which does not reveal if error is evenly distributed among classes or if some classes have very low accuracies and others have very high accuracies. In an imbalanced scenario, the total accuracy is not an appropriate measure to evaluate classification accuracy performance for very small minority classes since the accuracy may mainly represent the classification performance of the majority classes due to the dominance of the majority classes.

In this study, the citrus land cover classification problem is formulated as a binary classification: citrus or "other". The Florida citrus crop covers just around 3.2% of total land cover within the study area. It is a typical minority class, which is overwhelmed by the majority classes. The citrus class is the targeted class. In the binary classification confusion matrix, the targeted class is set as the positive class and the "other" class is designated as the negative class. To better measure classification accuracy, we propose a balanced accuracy (BA) measure of a targeted class, as defined as following:

$$Balanced\ \ Accuray(TCTA) = \frac{TP}{TP + FN + FP}$$

The balanced accuracy measure incorporates both False Positive and False Negative errors which directly affect the targeted class. It thus truly reflects the classification accuracy of the targeted class, not including the accuracy information of the "other" class.

In addition, the Kappa coefficient is also a popular measure used to assess the accuracy of the land cover classification. The Kappa coefficient reflects the difference between actual agreement and the agreement expected by chance. The estimate of Kappa coefficient $\hat{k}$ is defined as following

$$\hat{k} = \frac{N\sum_{i=1}^{c} x_{ii} - \sum_{i=1}^{c}(x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{c}(x_{i+} * x_{+i})}$$

where C and N represent the number of classes, and the total number of samples, respectively. x_{ii}, x_{i+} and x_{+i} represent the correctly classified pixel numbers in class *i*, the sum of the class *i* in the classified data, and the sum of class *i* in the validation data, respectively.

4. RESULTS AND DISCUSSIONS

The classification accuracy results with respect to the different training sampling ratios (TSR) are summarized in Table 2. The training sampling ratio of citrus to "other" class runs from 2%, 3%, 5%, 10%, 20%, 30%, 40% to 50%. As shown in Table 2, the producer's accuracy monotonically increases as the training sample ratio of citrus to "other" class increases while the user's accuracy decreases. The balanced targeted accuracy (BA) – citrus balanced accuracy starts at a lower 70.73% for 2% TSR, increases to the best accuracy of 71.41% for 3% TSR and then the accuracy monotonically decreases to the lowest 42.14% as TSR increases to a class balance ratio (50%). As shown in Table 2, the class balancing does improve the omission errors but significantly increases commission errors. It does not improve the targeted class balanced accuracy at all. As observed from Table 2, the best training sample ratio for a targeted citrus class should be in proportion to the ratio of the actual population sizes of classes. In addition, this result also indicates that the balanced accuracy measure is more appropriate for classification imbalance problem performance assessment.

Interestingly, the overall Kappa coefficient follows the same pattern of change as the balanced accuracy. This means it can also be used as criterion for accuracy performance improvement assessment.

5. CONCLUSIONS

This paper demonstrates experimentally that the training sample ratio of imbalanced classes significantly affects the classification accuracy using a See5 decision tree classifier for the class imbalance problem. It is found that balancing classes does not improve the classification accuracy of the minority class using a See5 decision tree classifier. The study results indicate that balancing classes by equalizing instances of classes will not improve performance. However, selecting a training sample ratio which reflects the actual ratio of the land cover classes will yield the best classification performance.

The proposed targeted class balanced accuracy (BA) measure provides a better measure of classification accuracy for a targeted individual class. It incorporates both False Positive and False Negative errors to truthfully reflect the accuracy of the targeted individual class. Moreover, the overall Kappa coefficient also truthfully reflects the changes in classification accuracy similar to the BA.

TABLE 2. Citrus accuracy with different minority/majority class sampling ratios

Training Sampling Ratio (TRS)	Citrus Producer Accuracy	Citrus User Accuracy	Overall Kappa	Citrus Balanced Accuracy (BA)
2%	85.3%	80.6%	0.823	70.73%
3%	89.1%	78.3%	0.828	71.41%
5%	91.9%	75.6%	0.824	70.88%
10%	93.9%	71.6%	0.806	68.41%
20%	95.4%	64.0%	0.757	62.07%
30%	96.4%	56.7%	0.702	55.49%
40%	97.2%	49.4%	0.640	48.70%
50%	97.8%	42.5%	0.575	42.14%

6. REFERENCES

[1] C. Boryan, Z. Yang, R. Mueller, and M. Craig, "Monitoring US Agriculture: The US Department of Agriculture, National Agricultural Statistics Service Cropland Data Layer Program," *Geocarto Int.* vol. 26, no. 5, pp. 341-358, 2011.

[2] G.M. Foody, and A. Mathur, "The Use of Small Training Sets Containing Mixed Pixels for Accurate Hard Image Classification: Training on Mixed Spectral Responses for Classification by a SVM," *Remote Sensing of Environment*, vol. 103, no. 2, pp. 179–189, 2006.

[3] N. Japkowicz, and S, Shaju, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429-449, 2002.

[4] X.Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Trans. on Systems, Man, and Cybernetics—Part B: Cybernetics*, 39(2), April 2009.

[5] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Intl. Conf. on Artificial Intelligence*, 2000.

[6] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," arXiv preprint arXiv:1710.05381, 2017.

[7] A. Ali, S. M Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A Review," *Int. J. Advance Soft Compu. Appl*, Vol. 7, No. 3, November 2015.

[8] D. Johnson, "Florida Commercial Citrus Inventory Now Maintained with GIS" Arc News Online, summer, 2006.

[9] C.G. Homer, et al., "Completion of the 2011 National Land Cover Database for the Conterminous United States-Representing a Decade of Land Cover Change information," *Photogramm. Eng. & Remote Sens*, 81(5), pp. 345-354, 2015.

[10] R.G. Congalton and K. Green, "*Assessing the Accuracy of Remotely Sensed Data -Principles and Practices*," Second edition (2009) CRC Press, Taylor & Francis Group, Boca Raton, FL 978-1-4200-5512-2.

Disclaimer: The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.