2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Coping with Imbalanced Training Data for Improved Terrain Prediction in Autonomous Outdoor Robot Navigation

Michael J. Procopio, Jane Mulligan, and Greg Grudic

***Abstract*— Autonomous robot navigation in unstructured outdoor environments is a challenging and largely unsolved area of active research. The navigation task requires identifying safe, traversable paths that allow the robot to progress towards a goal while avoiding obstacles. Machine learning techniques are well adapted to this task, accomplishing *near-to-far learning* by training appearance-based models using near-field stereo readings in order to predict safe terrain and obstacles in the far field. However, these methods are subject to degraded performance when training data sets exhibit *class imbalance*, or skew, where data instances of one class outnumber those in another. In such scenarios, classifiers can be overwhelmed by the *majority class*, and will tend to ignore the *minority class*. In this paper, we show that typical outdoor terrain scenarios are associated with training data imbalance, and examine the impact of using undersampling, oversampling, SMOTE, and biased penalties techniques to correct for imbalance in stereo-derived training data. We conduct a statistically significant, repeated measures empirical evaluation and demonstrate improved far-field terrain prediction performance when using such methods for handling class imbalance versus taking no corrective action at all.**

I. Introduction

Autonomous robot navigation in unstructured outdoor environments is a challenging and largely unsolved area of active research. The navigation task requires identifying safe, traversable paths that allow the robot to progress towards a goal while avoiding obstacles. Stereo vision allows for obstacle avoidance in the near field (here, within 10m of the robot). However, navigating solely on near-field terrain readings can lead to a common failure mode in outdoor autonomous navigation where incorrect trajectories are followed due to *nearsightedness*, or inability to distinguish safe and unsafe terrain the far field [1]. Being able to perceive safe terrain and obstacles in the far field allows for more natural (and efficient) paths to be planned and followed by the robot, while also aiding in the avoidance of *cul-de-sacs*.

To address near-sighted navigational errors, near-to-far learning is often used [2], [3], [4]. Framed as a supervised machine learning problem, the near-to-far approach (Sec. II-B) uses both appearance and stereo information from the near field as inputs for training appearance-based models, and then applies these models in the far field in order to predict safe terrain and obstacles farther out from the robot where stereo readings are unavailable (here, greater than 10m).

This work was conducted while the first author was a Senior Member of the Technical Staff at Sandia National Laboratories, Albuquerque, NM. mprocopio@gmail.com

Jane Mulligan is with the Department of Computer Science at the University of Colorado at Boulder, Boulder, CO, USA. janem@cs.colorado.edu

Greg Grudic is with Flashback Technologies, Longmont, CO, USA. Greg.Grudic@flashbacktechnologies.com

(a) LAGR Robot (b) Typical Scenario (c) Stereo Perception

Fig. 1: The LAGR robot (left); sample image from an outdoor scene (center); stereo obtained from onboard cameras (right).

In the example shown in Fig. 1, this approach could be used to predict the geometry of the far field, and hence obstacles and traversable terrain, beyond what stereo readings alone could provide. With such terrain predictions in the far field, the robot would follow a more natural path towards the goal, in this case avoiding trajectories towards far-field obstacles.

In general, supervised machine learning classifier design assumes when training a model that the distribution of the class labels in the training data is uniform, i.e., no *skew* or *class imbalance* is present in the training set. Often times, however, this is not the case, and classifier performance can degrade significantly as the severity of the imbalance increases. As a result, coping with imbalanced training data is the subject of ongoing research. Applications where class imbalance is a factor are increasingly common. Examples include detecting oil spills from satellite images [5], text classification [6], customer churn prediction [7], and so on; further examples are noted in the literature [8].

The autonomous outdoor robot navigation domain is also associated with and impacted by class imbalance. We demonstrate that typical terrain scenarios are associated with skew; in particular, the number of groundplane near-field stereo labels outnumbers the number of obstacle labels. In previous work [9], [10], [11], we observed reduced terrain prediction performance if specific action was not first taken to address this class imbalance. Moreover, alternative learning approaches described in related outdoor robot navigation research [12] are also susceptible to degraded classification due to class imbalance in the training data.

A. Research Objective and Contribution

This paper characterizes the class imbalance of near-field stereo labels in typical outdoor robot navigation scenarios, describes methods of coping with such skew, and reports empirical results conducted on natural datasets, quantifying the methods' performance benefit.

Towards this end, our research hypothesis is that far-field terrain prediction performance can be increased by handling training data skew, either by weighting minority class instances more heavily, or by creating balanced training data sets in various ways. To test this hypothesis, we conduct a statistically significant empirical evaluation using natural datasets taken from typical outdoor scenarios.

The contribution of this research is three-fold. In this paper, we:

1) Characterize and quantify the level of class imbalance present in near-field training data sets from typical outdoor robot navigation scenarios;
2) Contribute to the autonomous robot navigation literature by conducting a statistically significant experimental evaluation to determine the benefit of various methods for coping with class imbalance when doing near-to-far learning; and
3) Contribute to the class imbalance literature by examining the impact of skew correction when (a) using the logistic regression classifier, and (b) when the training data and test data distributions are different (i.e., near-field versus far-field terrain).

II. Background

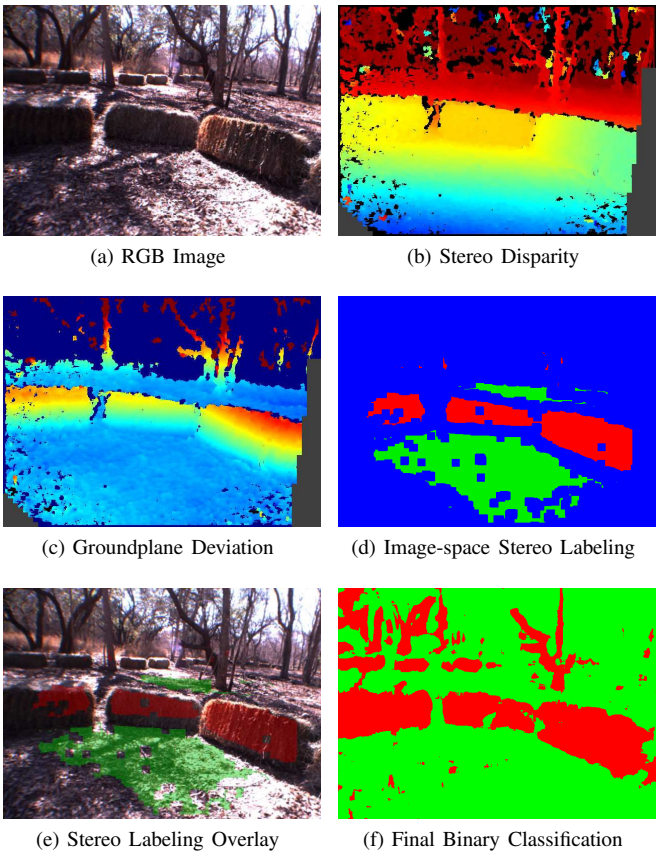
A. Related Work in Robotics and Vision

Approaches that use image appearance or color to segment regions of interest for navigation have existed since the 1980s [13], [14]. Research in autonomous robot navigation also has many decades of history and is ongoing [15], [16], [12]. More recently, programs such as DARPA's Learning Applied to Ground Robots (LAGR) program [1] have inspired work on using machine learning approaches to exploit image color and texture for classification of traversable terrain and obstacles in the far field [17], [18]. A more in-depth survey is given in [4].

B. Near-to-Far Learning Overview

Near-to-far learning using stereo is demonstrated in Fig. 2. For a given RGB image (2a), stereo disparity is computed using a stereo camera pair (2b). A groundplane model is fit and subtracted out, resulting in an estimate of groundplane deviation (2c). Near-field stereo labels from both the groundplane and obstacle classes are extracted according to small and large groundplane deviation values, respectively (2d); these near-field stereo labels form the training data set. Next, features are extracted from the image at the pixels in this stereo-labeled training set (2e); here, color histogram features are used [4]. A machine learning model is then trained on the resulting near-field feature data. The resulting model is evaluated over the remainder of the image, including the far field, to arrive at final terrain predictions (2f).

The model output is reconstructed to form a *cost image*, a pixelwise, image-space labeling of terrain with cost values, where high cost corresponds to obstacle and low cost corresponds to safe terrain. The cost image is then projected into the groundplane [19] creating a cost map used for robot navigation. The cost map is in turn sent to the planner [20], and in this manner, the low-level navigation of the robot is influenced. A more in-depth treatment is given in [4].



(a) RGB Image (b) Stereo Disparity (c) Groundplane Deviation (d) Image-space Stereo Labeling (e) Stereo Labeling Overlay (f) Final Binary Classification

Fig. 2: Demonstration of near-to-far learning using stereo. In (d), (e), and (f), red represents nontraversable obstacle (positive); green represents traversable groundplane (negative).

C. Base Learner: Logistic Regression

The classification algorithm for training and evaluating models in this research is *logistic regression* [21], a generalized linear classification method common in statistics and machine learning and appropriate for predicting the certainty of a binary outcome. Such models are very efficient to train on large-scale data, motivating their use in the real-time scenarios considered here. Moreover, these models by definition produce probabilistic output, which is desirable in robotic planning contexts. In this paper, we use the LIBLINEAR implementation for fast large-scale classification with logistic regression using a trust-region Newton method [22].

D. Prediction and Visualization Conventions

In this study, terrain predictions yielded by the logistic regression models are continuous on the interval $[0, 1]$. A predicted value of 0 represents full-confidence groundplane (negative) prediction. This is shown by green coloring. Similarly, a predicted value of 1 represents full-confidence obstacle (positive) prediction, shown by red coloring. In prediction images, color *intensity* correlates with prediction confidence. Thus, black represents full uncertainty (0.5), i.e., a test point on the decision boundary of the terrain model.

III. Stereo Label Class Imbalance

To demonstrate the degree to which class imbalance is present in terrain-based learning approaches, we characterize and quantify the degree of skew in stereo-derived near-field training data sets from typical outdoor scenarios. For this purpose, an analysis is performed using six hand-labeled natural data sets taken from the domain.

This data collection, recently contributed by the authors [10] and made publicly available [23], is referred to as the Hand-labeled DARPA LAGR datasets. Overall, three scenarios are considered; each scenario is associated with two distinct image sequences, representing different lighting conditions. Each of the six datasets consists of a 100-frame image sequence. The terrain in the images has been hand-labeled, with each pixel being placed into one of three classes: Obstacle, Groundplane, or Unknown. Fig. 3 gives a representative image from each dataset, while Fig. 7e shows an example of the ground truth labeling for a sample image from Dataset 1B. Further details are available in [4].

These datasets serve two purposes for this study. First, they are used to establish the presence and degree of skew in typical outdoor scenarios by examining associated near-field stereo information. Second, they provide the ground truth labels against which the classifiers, trained with various mechanisms for coping with skew, are evaluated.

Related to this, we note an important consideration in this research. Whereas typically, a training set is sampled from and is statistically reflective of the larger, general population, in this research, this is not necessarily the case. This is because distribution of the training data in the near-field can and does differ from that of the far-field.

Our analysis shows that stereo labels derived from the data above are associated with skew. This is illustrated by examining positive stereo label percentage over time (frame #) for each of six datasets, plotted in Fig. 4. Some datasets exhibit heavy skew, others have only mild skew, while some contain varying degrees of skew (as the terrain changes across the images). Corresponding summary statistics across each dataset are given in Table I. Examples from the datasets showing mild, moderate, and heavy skew are shown in Fig. 5.

(a) DS1A (b) DS1B (c) DS2A (d) DS2B (e) DS3A (f) DS3B

Fig. 3: Representative images from the six datasets in the Hand-Labeled DARPA LAGR Data [23]. Two lighting conditions each from three different terrain scenarios are included. Each sequence comprises 100 labeled frames.

IV. Coping with Imbalanced Training Data

The class imbalance problem occurs when training data comprises many more data instances from one class than another; in these situations, standard machine learning classifiers can be overwhelmed by the *majority class*, and will tend to ignore the *minority class*. The resulting models will have higher misclassification rates on the minority class [24].

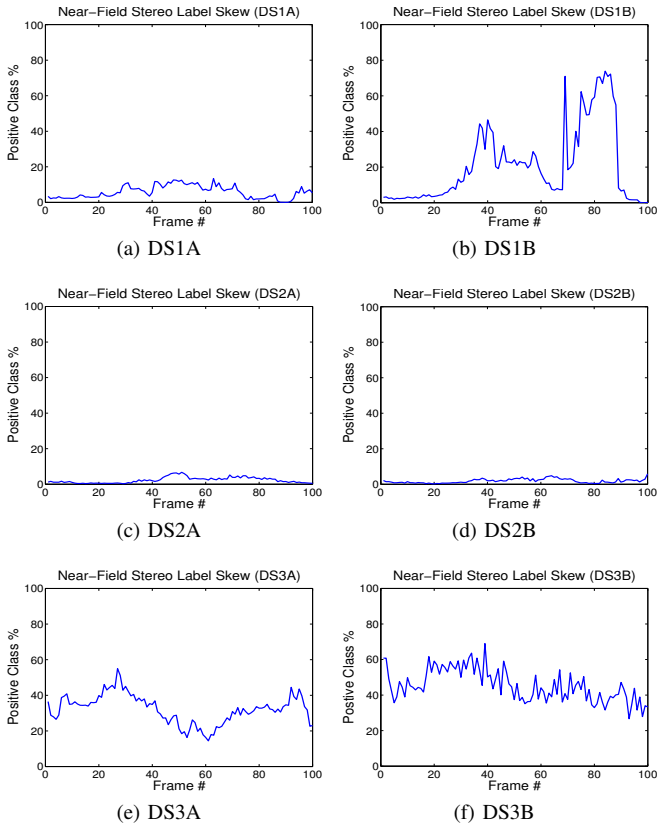
The problem of class imbalance in training data can be addressed in many ways; see [25] for an in-depth survey. For this study, we consider two widely accepted and commonly used general approaches. The first approach includes *sampling* or *sample balance* methods that operate at the data level in an algorithm-independent way [26]. The second approach includes classifier-specific methods that involve adjustments at the algorithm level during training time.

In addition, there are also other more specialized methods not considered in this study. For example, the field of active learning [27] provides methods for choosing training data; such methods are most classically applied when acquiring training labels is expensive, which is not the case here. Moreover, the real-time domain requirement for robot navigation precludes many active learning techniques, although recent work in this area proposes an efficient method for active learning in the presence of class imbalance [28].

A. Sample Balance Methods for Skew Correction

This general class of method seeks to achieve balanced training data sets by using a variety of data processing techniques, resulting in balanced sets that differ in the total number of instances versus the original.

Random Undersampling. The majority class is randomly undersampled so that the number of examples in this class is made to equal the number of examples in the minority class, yielding a balanced training set (Fig. 6d). Also known as one-sided sampling [29], undersampling can have computational advantages because the training set size is reduced. One objection to undersampling is that it ignores a certain amount of potentially useful majority-class examples [25].



(a) DS1A (b) DS1B (c) DS2A (d) DS2B (e) DS3A (f) DS3B

Fig. 4: Class imbalance in stereo-derived near-field training labels from the Hand-Labeled DARPA LAGR Data. DS1B has moderate skew; DS1A, DS2A, and DS2B have heavy skew; while DS3A and DS3B have only mild skew.

Random Oversampling. The minority class is randomly oversampled, with replacement, so that the number of examples in this class is made to equal the number of examples in the majority class, yielding a balanced training set (Fig. 6a). (In the figure, the training sets are in fact balanced, although they do not appear so; this is because the oversampled set contains *duplicates* of the original data points.) One objection to oversampling is that it can increase training time for a classifier, and in some situations, can lead to overfitting [25].

SMOTE. In contrast to random resampling (above), SMOTE (for *Synthetic Minority Oversampling TEchnique*) synthesizes *new* data from the minority class by interpolating new samples among a given minority class point's k nearest neighbors [30]; an example is shown in Fig. 6c. This technique belongs to a broader class of "informed" (versus merely random) sampling methods. Because of the synthesized minority class examples, the shape of the learned decision boundary can be made to be smoother.

Other Sampling Methods. Other more sophisticated sampling methods include *directed* approaches that oversample or undersample in an informed manner [24], instead of entirely at random. For example, when taking an oversampling approach, instead of sampling entirely at random, resampling could occur more frequently on training data closer to the decision boundary. Ensemble approaches have also been described [31] [32].

TABLE I: Stereo Label Statistics for DARPA LAGR Data

Number of stereo-labeled pixels, mean across all 100 frames by dataset

DATASET	TOTAL	NEG	POS	POS %	SKEW
DS1A	55,791	52,845	2,946	5.8 %	18 : 1
DS1B	61,825	50,808	11,017	20.9 %	5 : 1
DS2A	83,904	82,090	1,814	2.3 %	45 : 1
DS2B	78,526	77,148	1,377	1.9 %	56 : 1
DS3A	28,380	18,865	9,514	32.2 %	2 : 1
DS3B	11,089	5,841	5,247	45.4 %	1.1 : 1

B. Algorithmic Methods for Skew Correction

This general class of method seeks to counter class imbalance by any number of algorithm-specific methods that operate beyond the data level, i.e., internal to the actual classifier. One approach that is common in the literature, for classifiers that support it, is to adjust the misclassification cost more heavily for the underrepresented minority class [33]. This *biased penalties* method aims to ensure that minority-class instances have adequate influence over the decision boundary at training time [34]; this is shown in Fig. 6b. Classifiers that support different penalty factors for different classes include the Support Vector Machine (SVM) and logistic regression [22]; the latter is used in this research.

C. Computational Efficiency Considerations

Computational analysis of the skew correction methods can be divided into two parts: the cost for performing the correction, and the resultant impact on model training. Consider a scenario with N and J samples in the minority and majority classes, respectively.

For random undersampling, $J-N$ majority class samples are removed at random, and instead of training a model using all $J+N$ samples, a classifier is now trained with $2N$ samples. This can result in significantly reduced training time. For random oversampling, $J-N$ minority class samples are added (duplicated) at random, and a model is trained using $2J$ samples; this generally results in increased training time.

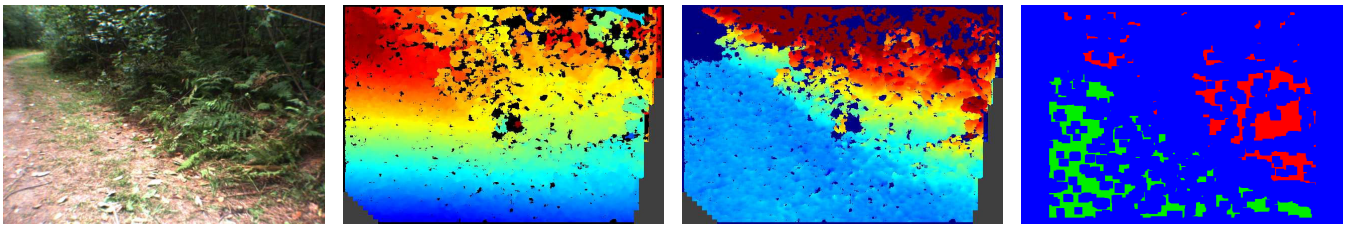
SMOTE is more expensive. Nominally (for balanced sets of approximately equal size), $J-N$ samples must be synthesized. This requires interpolation between k nearest neighbors for some number of minority class examples, which can be computationally intensive. The resulting larger set will also generally lead to increased training time.

Finally, for the biased penalties approach, class weights must be computed; the actual training data set is not modified. The additional computational cost of using class-specific weights when training models will be classifier-dependent; for SVM and logistic regression, it is negligible.

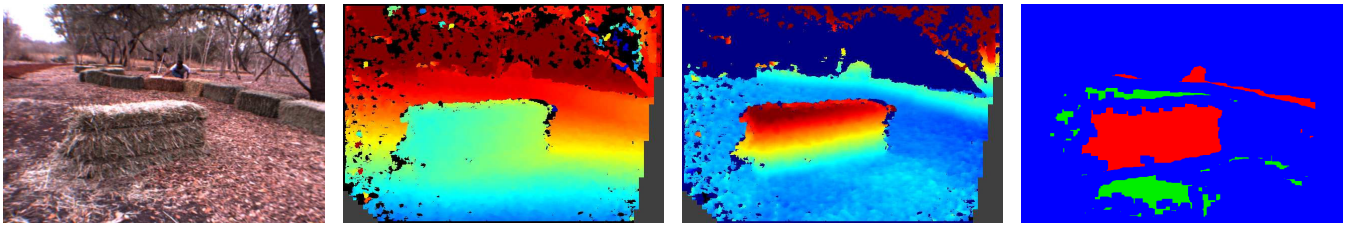
V. EXPERIMENTS

A. Experimental Approach

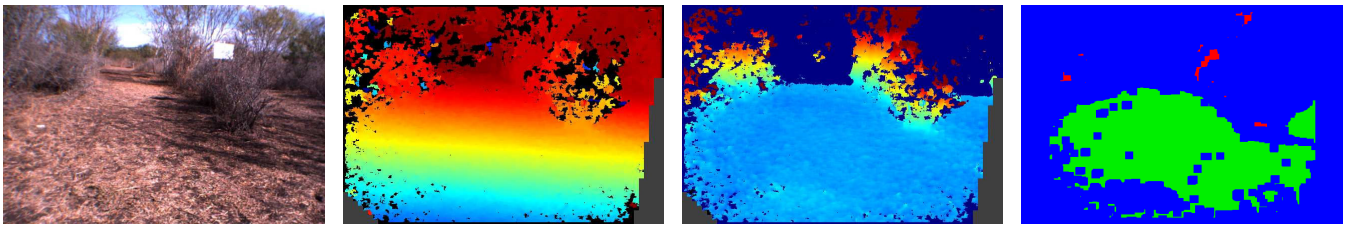
Hypothesis. Our research hypothesis is that far-field terrain prediction performance can be increased through handling skew in training data sets, either by influencing the decision boundary with the biased penalties approach or by sampling training data to create balanced training sets.



(a) RGB Image – DS3A Frame 26 (b) Stereo Disparity (c) Groundplane Deviation (d) Stereo Mask (Light Skew)



(e) RGB Image – DS1B Frame 82 (f) Stereo Disparity (g) Groundplane Deviation (h) Stereo Mask (Moderate Skew)



(i) RGB Image – DS2B Frame 34 (j) Stereo Disparity (k) Groundplane Deviation (l) Stereo Mask (Heavy Skew)

Fig. 5: Examples of mild, moderate, and heavy skew found in the Hand-Labeled DARPA LAGR Data collection.

Method. To test this hypothesis, and to evaluate the impact of the various methods for handling class imbalance in this domain, we conducted randomized experiments on log data from test runs conducted during the LAGR program in typical, but challenging, outdoor terrain scenarios under a variety of lighting conditions. A repeated measures design is needed to evaluate the statistical significance of the sample balancing methods, whose results will naturally vary over individual experiments due to their associated random component. For this study, we ran 10 randomized experiments.

Testing Sequence. For each image in each dataset, a single model is learned from training data extracted from near-field stereo labels associated with that image. This model is then applied to the pixels in that entire image, yielding probabilistic terrain class predictions, and then discarded. This is done separately for each of the four skew correction methods in the study, as well as for the baseline method (no skew correction).

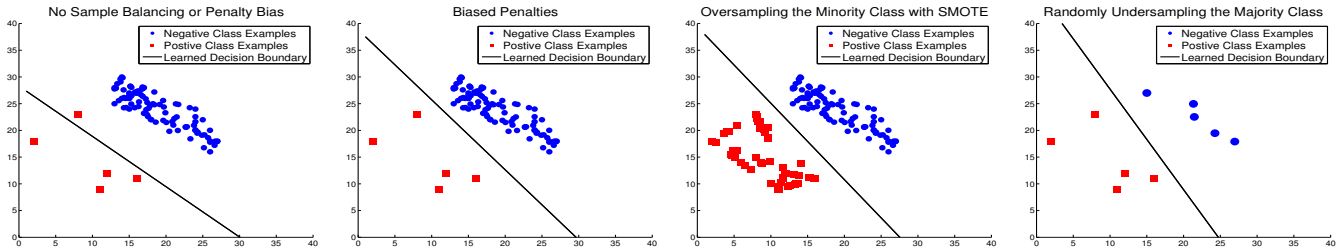
For example, the undersampling approach will create a balanced training set prior to training the model, while the biased penalties approach will use all of the training data and influence the model using cost parameters proportional to the skew. Finally, this procedure is repeated 10 times, yielding 10 randomized experiments in total; this is needed due to the randomness inherent in the sample balance methods.

Evaluation. To evaluate classifier performance, probabilistic classifier outputs (terrain predictions) are scored against discrete ground truth labelings included in the hand-labeled DARPA LAGR data. Our primary aim is to identify safe terrain and obstacles in the far field (since traditional stereo approaches are generally able to identify obstacles in the near field). Therefore, we only score algorithm output on pixels in the far field (approximately 10m out from but within 100m of the robot). Additional considerations regarding use of the *far-field band* for evaluation purposes are given in [4].

The final score as reported in Table II is the average of the performance of the algorithm over all 100 images in the dataset. In particular, the score for a particular image is given by computing the pixelwise RMSE (see below) of the probabilistic terrain predictions in the far field as compared to the discrete class labels from the ground truth.

Fixed Parameters. Some of the methods in this study have parameters whose values must be specified:

1) For SMOTE, we set k (the number of nearest neighbor points from which to interpolate when synthesizing data) to 5 as done by the algorithm designers in their initial research [30]. The oversampling (SMOTE) percentage is data-driven and is set to whatever value is needed (in integral multiples of 100, as required by SMOTE) for the size of the minority class to approximately match the majority class, resulting in generally balanced sets with negligible skew.
2) For the biased penalties approach, the values for the two cost parameters C_{neg} and C_{pos} are determined directly from the degree of skew in the training data (see [33]). In particular, these parameters are proportional to the degree of skew, and the values are scaled such that $C_{neg} + C_{pos} = 1$.



(a) No Skew Adjustment (b) Biased Penalties (c) Oversampling with SMOTE (d) Random Undersampling

Fig. 6: Impact of various methods for coping with imbalanced training data on the learned decision boundary.

Performance Metric. The performance metric used in this evaluation is Root Mean Square Error (RMSE), where lower scores are better. RMSE is applicable to binary classification scenarios where the classifier predictions are continuous on $[0, 1]$ for corresponding discrete target labels in $\{0, 1\}$:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2} \quad (1)$$

where $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_N$ are the probabilistic classifier predictions on $[0, 1]$ for a set of N test points, and $y_1, y_2, \ldots, y_N$ are the corresponding discrete class labels in $\{0, 1\}$.

RMSE is particularly useful for this scenario because no specific threshold must be defined, e.g., as would be done for binary classification accuracy. Importantly, the use of RMSE here penalizes higher confidence incorrect predictions (i.e., those prediction probabilities approaching 0 or 1) *more so* than it does for lower confidence predictions (i.e., those predictions approaching 0.5). In this sense, RMSE as applied here is a generalized and continuous scoring mechanism with no discontinuity at an arbitrary threshold. Additional details and rationale for the use of RMSE are given in [4].

B. Experimental Results and Discussion

Raw experimental data from the study is given in Table II. Overall, these data support two key conclusions. First, it is clear that methods for handling class imbalance—including the three sampling methods and the algorithmic biased penalties approach—have a significant, positive impact on performance versus taking no action to address the skew. Second, as a group, the three sample balance methods all outperformed the biased penalties algorithmic approach.

Over all six datasets, oversampling performed the best (lowest mean RMSE score), followed by undersampling, SMOTE, and then biased penalties. These results were all significant at the 95% confidence level using the unmatched pairs *t*-test (where the scores from the randomized experiments are the independent samples).

With regard to the statistical analysis, we emphasize that statistical significance was determined by comparing the mean and variance of the scores from the 10 repeated measures (the randomized experiments). This is in contrast to achieving statistical significance using high sample sizes derived from the large number of pixels in an image.

Within each dataset, there was not always a statistically significant difference among the three sampling methods. The three sampling methods did outperform the biased penalties approach for all datasets except DS1A. The biased penalties approach outperformed the baseline (no skew correction) for all datasets except DS1B. Within each dataset, and overall, all three sampling methods performed better versus taking no skew correction action at all.

A final important observation we make is that performance benefit (reduction in error) of using sampling methods versus taking no corrective action for skew was generally linear with the average degree of skew present in the data. This fits the intuition that classifier performance degrades as the severity of class imbalance increases.

Based on the above empirical findings, we offer the following guidance for the community. First, if it is the case that there is class imbalance in the training data (stereo-derived or otherwise), some type of action to correct for skew should be taken for optimal performance. This is a result echoed often in the class imbalance applications literature. Moreover, our findings suggest that, for optimal prediction performance, oversampling is the method of choice. If this approach precludes real-time performance, undersampling should be considered.

C. Experimental Snapshots and Narrative

Representative output from four scenarios is shown in Fig. 7. The first scenario (Fig. 7a) is associated with approximately 4:1 skew (7b). Although there is generally enough training information from both classes to make reasonable terrain predictions without any balancing (7c), by using undersampling to achieve a balanced training set, obstacle predictions in the far-field are more robust (7d). Note, however, that false-positive obstacle predictions (lower-left) are also present to a larger degree.

The second and third scenarios (Figs. 7f and 7k) are associated with heavier skew—around 40:1. In these scenarios, entire patches of mid- and far-field obstacles (dense, leafless foliage on the right) are not detected, because they are represented only minimally in the stereo-derived training sets. Using the biased penalties approach (7i) or oversampling (7n), these obstacles are adequately identified and the robot's trajectory would be adjusted accordingly.

The fourth scenario (Fig. 7p) is associated with milder skew, around 2:1 (7q). Terrain predictions without skew correction are reasonable (7r). By oversampling using SMOTE, terrain predictions are more robust; obstacle predictions are confident and obstacle regions are more contiguous (7s).

TABLE II: Summary of Experimental Results – RMSE

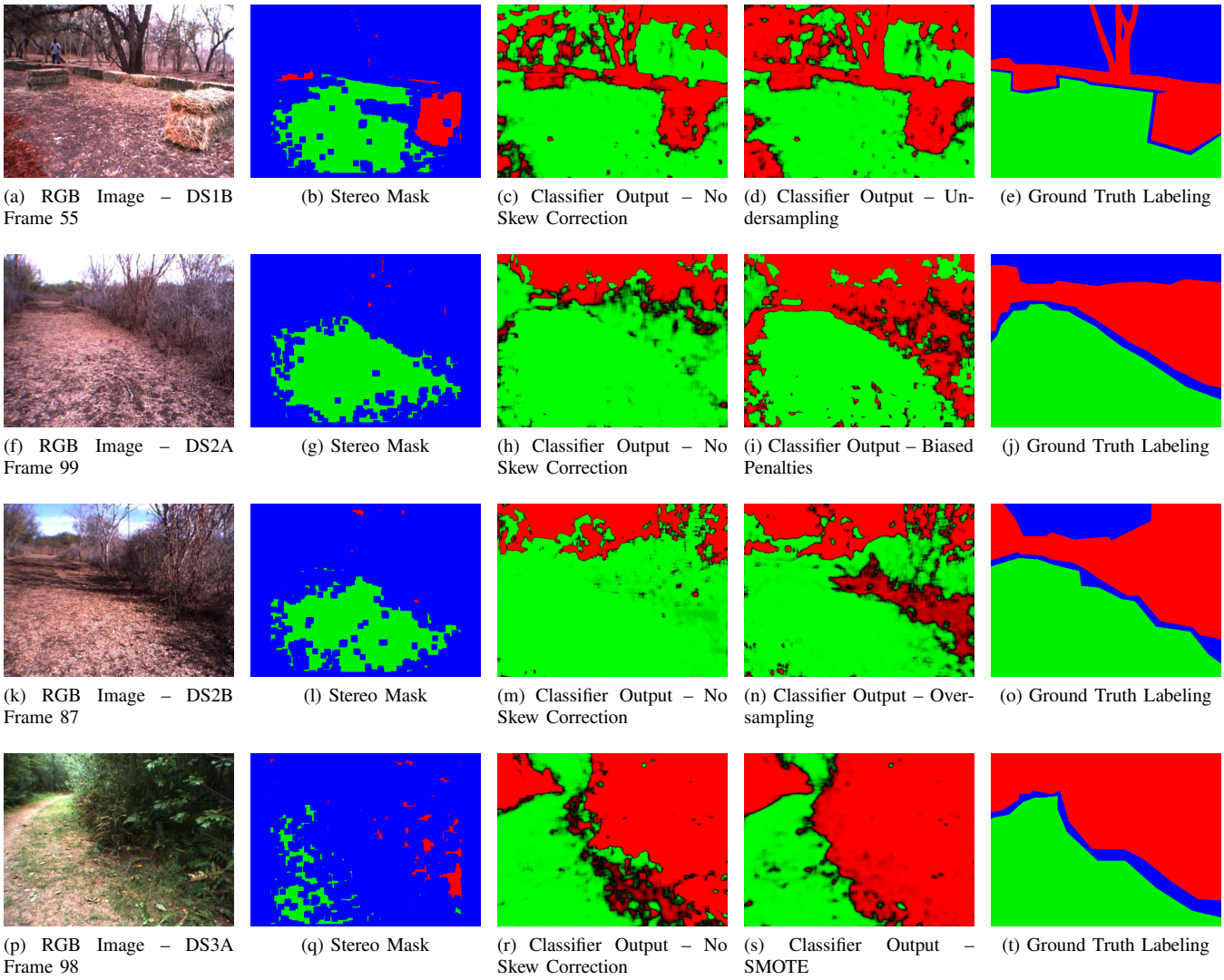
Mean framewise Root Mean Square Error (RMSE) over entire dataset – ***lower scores are better***

DATASET	NO ADJUSTMENT[a]	BIASED PENALTIES[a]	UNDERSAMPLING[b]	OVERSAMPLING[b]	SMOTE[b]
DS1A	0.486	0.416	0.424 ±0.002	0.423 ±0.001	0.423 ±0.000
DS1B	0.272	0.278	0.263 ±0.005	0.257 ±0.003	0.263 ±0.002
DS2A	0.393	0.261	0.221 ±0.002	0.220 ±0.001	0.224 ±0.000
DS2B	0.676	0.505	0.500 ±0.001	0.498 ±0.000	0.502 ±0.000
DS3A	0.104	0.100	0.097 ±0.000	0.097 ±0.000	0.100 ±0.000
DS3B	0.139	0.137	0.134 ±0.001	0.134 ±0.001	0.135 ±0.000
OVERALL[c]	0.345	0.283	0.273 ±0.0008	0.271 ±0.0006	0.274 ±0.0004

[a] No random component in method, hence no variance reported among randomized experiments.

[b] Standard deviation of 10 repeated measures (randomized experiments).

[c] Overall performance, mean over all datasets.



(a) RGB Image – DS1B Frame 55 (b) Stereo Mask (c) Classifier Output – No Skew Correction (d) Classifier Output – Undersampling (e) Ground Truth Labeling

(f) RGB Image – DS2A Frame 99 (g) Stereo Mask (h) Classifier Output – No Skew Correction (i) Classifier Output – Biased Penalties (j) Ground Truth Labeling

(k) RGB Image – DS2B Frame 87 (l) Stereo Mask (m) Classifier Output – No Skew Correction (n) Classifier Output – Oversampling (o) Ground Truth Labeling

(p) RGB Image – DS3A Frame 98 (q) Stereo Mask (r) Classifier Output – No Skew Correction (s) Classifier Output – SMOTE (t) Ground Truth Labeling

Fig. 7: Sample experimental snapshots from the study. Explanation and narrative is given in Sec. V-C. Red color indicates obstacle (nontraversable) terrain prediction, green color indicates ground plane (traversable) terrain prediction, and color intensity indicates prediction confidence, with black representing full uncertainty (see Sec. II-D). For the ground truth images at far right, blue color indicates unknown areas in the ground truth; these unlabeled regions are not used in the computation of classifier performance.

VI. Conclusions and Future Work

This paper forms part of a line of research that examines the use of supervised machine learning methods to predict terrain class, i.e., obstacle or groundplane, in autonomous outdoor robot navigation. This paper examined the impact of *class imbalance*, or skew, in training data sets on terrain prediction performance, with the hypothesis that taking specific action to correct for class imbalance will lead to increased terrain prediction performance in the far field. A statistically significant empirical evaluation was conducted on natural, hand-labeled ground truth datasets previously logged during outdoor robot navigation test runs.

The key empirical contributions of the paper are threefold. First, it was shown that typical outdoor scenarios are associated with varying degrees of skew in the stereo labels, with some scenarios having heavy skew (greater than 50:1). Second, it was shown that coping with this training data imbalance is critical to achieving optimal far-field terrain prediction performance. Finally, the experimental results indicated that overall, skew correction using *sample balance* methods such as random undersampling, random oversampling, and SMOTE outperformed the *biased penalties* approach, which in turn outperformed taking no action at all; these were all statistically significant results.

Future work could investigate the benefit of more sophisticated sampling approaches for skew correction. First, beyond basic random undersampling/oversampling, so-called *directed* sampling methods could be used [24]. A hybrid approach, which combines sample balancing with biased penalties [35], could also improve performance.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of Sandia National Laboratories, the National Science Foundation, the DARPA LAGR program, Philip Kegelmeyer, Chih-Jen Lin, and the anonymous reviewers' insightful comments.

References

[1] L. Jackel, E. Krotkov, M. Perschbacher, J. Pippine, and C. Sullivan, "The DARPA LAGR program: Goals, challenges, methodology, and Phase I results," *Journal of Field Robotics*, vol. 23, pp. 945–973, November/December 2006.

[2] A. Howard, M. Turmon, L. Matthies, B. Tang, A. Angelova, and E. Mjolsness, "Towards learned traversability for robot navigation: From underfoot to the far field," *Journal of Field Robotics*, vol. 23, pp. 1005–1017, November/December 2006.

[3] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski, "Self-supervised monocular road detection in desert terrain," in *Proceedings of Robotics: Science and Systems*, 2006.

[4] M. J. Procopio, J. Mulligan, and G. Grudic, "Learning terrain segmentation with classifier ensembles for autonomous robot navigation in unstructured environments," *Journal of Field Robotics*, vol. 26, no. 2, pp. 145–175, 2009.

[5] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, no. 2-3, pp. 195–215, 1998.

[6] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 80–89, 2004.

[7] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3, 2009.

[8] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, 2002.

[9] M. J. Procopio, J. Mulligan, and G. Grudic, "Long-term learning using multiple models for outdoor autonomous robot navigation," in *IEEE/RSJ Int'l Conf. on Intel. Robots and Systems*, pp. 3158–3165.

[10] ——, "Learning in dynamic environments with *Ensemble Selection* for autonomous outdoor robot navigation," in *IEEE/RSJ International Conf. on Intelligent Robots and Systems (IROS)*, 2008, pp. 620–627.

[11] M. J. Procopio, W. P. Kegelmeyer, G. Grudic, and J. Mulligan, "Terrain segmentation with on-line mixtures of experts for autonomous robot navigation," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science (LNCS), vol. 5519, June 2009, pp. 385–397.

[12] P. Bellutta, L. Matthies, K. Owens, and A. Rankin, "Terrain perception for DEMO III," in *In Proceedings of the IEEE Intelligent Vehicles Symposium 2000*, 2000, pp. 326–331.

[13] J. Crisman and C. Thorpe, "Unscarf, a color vision system for the detection of unstructured roads," in *Proc. 1991 Int. Conf. on Robotics and Automation*, Sacramento, CA, April 1991, pp. 2496–2501.

[14] D. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Adv. in Neural Information Proc. Systems 1*, 1989.

[15] T. Kanade, C. Thorpe, and W. Whittaker, "Autonomous land vehicle project at CMU," in *CSC '86: Proceedings of the 1986 ACM fourteenth annual conference on Computer science*, 1986, pp. 71–80.

[16] A. Kelly and A. Stentz, "An analysis of requirements for rough terrain autonomous mobility," *Auton. Robots*, vol. 4, no. 4, 1997.

[17] M. Happold, M. Ollis, and N. Johnson, "Enhancing supervised terrain classification with predictive unsupervised learning," in *Proceedings of Robotics: Science and Systems*, 2006.

[18] G. Grudic and J. Mulligan, "Outdoor path labeling using polynomial mahalanobis distance," in *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, August 2006.

[19] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall, 2003.

[20] J.-C. Latombe, *Robot Motion Planning*. Boston, Mass: Kluwer Academic Publishers, 1991.

[21] D. R. Cox and E. J. Snell, *Analysis of Binary Data*, 2nd ed. London: Chapman & Hall, 1989.

[22] C.-J. Lin, R. C. Weng, and S. S. Keerthi, "Trust region Newton method for large-scale logistic regression," *J. Mach. Learn. Res.*, vol. 9, pp. 627–650, 2008.

[23] M. J. Procopio, "Hand-labeled DARPA LAGR datasets," Available at http://www.mikeprocopio.com/research/labeledlagrdata/, 2007.

[24] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[25] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.

[26] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.

[27] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, 2002.

[28] S. Ertekin, J. Huang, and C. L. Giles, "Active learning for class imbalance problem," in *SIGIR '07: Proc. of the 30th International ACM SIGIR Conf. on R&D in Info. Retrieval*, 2007, pp. 823–824.

[29] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *In Proc. of the Fourteenth Int'l Conference on Machine Learning (ICML)*, 1997, pp. 179–186.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res. (JAIR)*, vol. 16, pp. 321–357, 2002.

[31] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory under-sampling for class-imbalance learning," in *ICDM '06: Proc. of the Sixth International Conf. on Data Mining*. IEEE, 2006, pp. 965–969.

[32] M. A. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "A multiple expert approach to the class imbalance problem using inverse random under sampling," in *Multiple Classifier Systems*, 2009, pp. 82–91.

[33] B. Raskutti and A. Kowalczyk, "Extreme re-balancing for svms: a case study," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 60–69, 2004.

[34] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *Workshop on Learning from Imbalanced Data Sets in International Conference on Machine Learning (ICML)*.

[35] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *In Proceedings of the 15th European Conference on Machine Learning (ECML)*, 2004, pp. 39–50.