HS-, HiS-, HeRS-Boot: Heterogeneity-Stratified Bootstrap, Enhancing Substance Recognition for Assistive technologies in Real-World Applications

Pertami J. Kunz and Syrine ben Abid
pertami.kunz@ieee.org
syrinebenabid@ieee.org
Technische Universität Darmstadt
Germany

ABSTRACT

We introduce the Heterogeneity-Stratified, the improved Heterogeneity-Stratified, and the Heterogeneity Ratio-Stratified Bootstrap oversampling methods (HS-, HiS, and HeRS-Boot, respectively), which assign higher resampling probabilities to sample points in less homogeneous regions. We demonstrate its advantage in the case of training a detector by oversampling the under-represented class in an imbalanced data set. We took a case study of spoiled food and allergen detectors in form of an electronic nose. Results demonstrate the effectiveness and generalizability of our method across different sensors, highlighting its potential for real-world applications and positive impact on daily life.

CCS CONCEPTS

• **Mathematics of computing** → **Bootstrapping**.

KEYWORDS

Oversampling, Bootstrap, Out of Bag Bootstrap, Detection, Classification, IoT

ACM Reference Format:
Pertami J. Kunz and Syrine ben Abid. 2023. HS-, HiS-, HeRS-Boot: Heterogeneity-Stratified Bootstrap, Enhancing Substance Recognition for Assistive technologies in Real-World Applications. In *Proceedings of WomENcourage ACM 2023*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

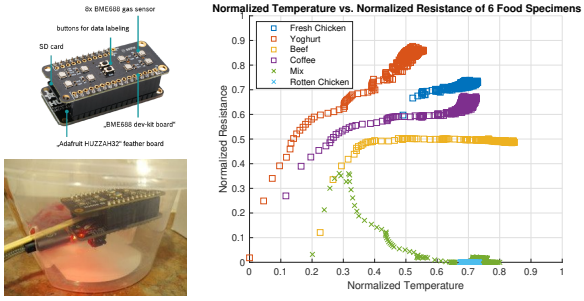
Anosmic or visually impaired people may fail to recognise spoiled foods or allergend in their fridge or pantry. A smart detector that can be trained to recognise such hazards is needed. Intelligently choosing a method to train the algorithm may reduce dependencies on the complexity and the amount of data collected. It is often the case that the class distribution in a dataset is not equal. A balanced training dataset may be created with an artificially equal class distribution by oversampling the instances in the minority class. In this paper, we introduce three oversampling methods based on the stratified resampling with replacement, or stratified bootstrapping.

2 METHODOLOGY

For the dataset, the gas compositions of the following specimens were measured The BME688 (Bosch Sensortec) sensor: (1) Fresh Chicken: a piece of fresh, raw chicken, (2) Yoghurt: fresh yoghurt (3) Beef: a piece of fresh raw beef, (4) Coffee: a handful of coffee beans (5) Mix : a piece of spoiled raw chicken, mixed with some fresh vegetables (6) Rotten Chicken : a piece of spoiled raw chicken. The measurement setup and the plot of the measurements are shown in Figure 1. The proposed algorithms are summarised in Algorithm 1.

Algorithm 1 Oversampling with the HS-, HiS, or HeRS-Boot [1, 2]

Step 1 Divide the sample domain into K grids.
Step 2 For each grid k, $k = 1, \ldots, K$, calculate the heterogeneity measure $H(C, k)$ and modify the resampling probability distribution $p_c(k)$ for each class c in cluster k (Table 1).
Step 3 Oversampling: Sample N_0 instances with replacement from Class 0 (non-target) and likewise N_0 instances from Class 1 (target) with the updated distribution p'_c.



(a) E-nose

(b) Measurements plot

Figure 1: Training the electronic nose with different specimens.

3 RESULTS AND DISCUSSION

Tables 2 and 3 show the detection metrics averaged over 500 trials, where TP is the true positives, TN the true negatives, FP the false positive, and FN the false negatives: (1) **Sensitivity (SN)** or true positive rate (TPR) or recall (REC) = TP / (TP + FN), (2) **Specificity (SP)** or true negative rate (TNR) = TN / (TN + FP), (3) **Accuracy (ACC)** = (TP + TN) / (TP + TN + FN + FP), (4) **Error rate (ERR)** =

WomENcourage ACM 2023, Sept 20–22, 2023, Trondheim, Norway

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Table 1: Summary of Algorithms, where γ is a hyperparameter constant that determines the influence of the heterogeneity, a_{ck} is the number of samples in grid k that belongs to class c, and the denominator is to make sure that the sum equals 1.

HSBoot, HiSBoot	HeRSBoot
$H(C,k) = -\sum_{c=1}^{C} \frac{a_{ck}}{N} \log\left(\frac{a_{c,k}}{\sum_1^C a_{c,k}}\right)$	$H(C,k) = -\sum_{c=1}^{C} \frac{\rho_{ck}}{K} \log\left(\rho_{c,k}\right)$
HSBoot	**HiSBoot, HeRsBoot**
$p'_c(k) = \frac{\frac{1}{N_c}+\gamma H(C,k)}{1+\gamma \sum_k a_{ck} H(C,k)}$	$p'_c(k) = \begin{cases} \frac{\frac{1}{N_c}+\gamma H(C,k)}{1+\gamma \sum_k a_{ck} H(C,k)} & \text{for target class} \\ p_c(k) & \text{for non target class} \end{cases}$

(FP + FN) / (TP + TN + FN + FP), (5) **Precision (PREC)** = TP / (TP + FP), and (6) **F1 score (F1)** = 2 * PREC * REC / (PREC + REC).

The Naive Bayes, Quadratic Discriminant, and Neural Network algorithms were used as classifiers, where in each trial, the hyperparameters were decided with leave 0.3 out cross validation. We can see from the tables, the proposed methods with different λ improve the traditional training set selection method.

Table 2: Spoiled Food Detection Results

(a) Naive Bayes

	Random	HiSBoot^2,			HeRSBoot^2		
		$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
SN	.9899	.9885	.9922	.9902	.9911	**.9924**	**.9924**
SP	.9924	**.9953**	.9945	.9927	.9946	.9923	.9917
ACC	.9916	.9930	**.9937**	.9919	.9934	.9923	.9920
ERR	.0084	.0070	**.0063**	.0081	.0066	.0077	.0080
PREC	.9850	**.9906**	.9891	.9856	.9892	.9848	.9837
F1	.9872	.9894	**.9906**	.9877	.9901	.9885	.9879

(b) Quadratic Discriminant

	Random	HiSBoot^2,			HeRSBoot^2		
		$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
SN	.9941	.9916	.9931	**.9953**	.9925	.9944	.9948
SP	.9913	**.9947**	.9930	.9901	.9938	.9911	.9917
ACC	.9922	**.9937**	.9930	.9918	.9933	.9922	.9928
ERR	.0078	**.0063**	.0070	.0082	.0067	.0078	.0072
PREC	.9831	**.9898**	.9863	.9807	.9879	.9826	.9837
F1	.9885	**.9906**	.9896	.9879	.9901	.9884	.9892

(c) Neural Network with 2 hidden layers, with 8 and 4 fully connected outputs for each hidden layer, respectively.

	Random	HiSBoot^2,			HeRSBoot^2		
		$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
SN	**.9869**	.9303	.9446	.9309	.9208	.9531	.9352
SP	.9253	.9654	**.9664**	.9569	.9605	.9535	.9518
ACC	.9459	.9537	**.9591**	.9482	.9473	.9534	.9462
ERR	.0541	.0463	**.0409**	.0518	.0527	.0466	.0538
PREC	.9094	.9406	**.9429**	.9274	.9325	.9244	.9194
F1	.9449	.9652	**.9667**	.9579	.9607	.9562	.9536

Table 3: Allergen Detection Results

(a) Naive Bayes

	Random	HiSBoot^2,			HeRSBoot^2		
		$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
SN	.6053	.7992	.7583	.7128	.8122	**.8157**	.7823
SP	.9810	.9762	.9793	**.9816**	.9744	.9728	.9735
ACC	.9184	.9467	.9425	.9368	**.9474**	.9466	.9417
ERR	.0816	.0533	.0575	.0632	**.0526**	.0534	.0583
PREC	.8655	.8723	.8819	**.8877**	.8653	.8590	.8576
F1	.7020	.8296	.8117	.7870	**.8338**	.8325	.8140

(b) Quadratic Discriminant

	Random	HiSBoot^2,			HeRSBoot^2		
		$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
SN	.8057	.7591	.7931	.7509	.7503	.8310	**.8474**
SP	.9653	.9701	.9697	**.9762**	.9704	.9653	.9631
ACC	.9387	.9349	.9402	.9387	.9337	.9429	**.9438**
ERR	.0613	.0651	.0598	.0613	.0663	.0571	**.0562**
PREC	.8228	.8353	.8401	**.8647**	.8353	.8273	.8211
F1	.8141	.7952	.8154	.8030	.7904	.8289	**.8334**

(c) Neural Network with 2 hidden layers, with 8 and 4 fully connected outputs for each hidden layer, respectively.

	Random	HiSBoot^2,			HeRSBoot^2		
		$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 5$
SN	**.7252**	.6395	.6399	.6276	.6228	.6506	.6474
SP	.9620	.9805	.9811	**.9824**	.9801	.9768	.9764
ACC	.9225	.9237	**.9242**	.9233	.9205	.9225	.9216
ERR	.0775	.0763	**.0758**	.0767	.0795	.0775	.0784
PREC	.8315	.8681	.8707	**.8782**	.8627	.8496	.8465
F1	**.7737**	.7548	.7656	.7690	.7524	.7653	.7702

SELECTED REFERENCES

[1] Pertami J. Kunz and Abdelhak M. Zoubir. 2023. Heterogeneity-Stratified Bootstrap Oversampling for Training a Spoiled Food Detector. (2023). Submitted to DSP 2023, Rhodos, Greece.

[2] Pertami J. Kunz and Abdelhak M. Zoubir. 2023. The improved Heterogeneity and Heterogeneity Ratio-Stratified Bootstrap Oversampling for Training an Allergen Detector. (2023). Draft.