Boundary bagging to address training data issues in ensemble classification

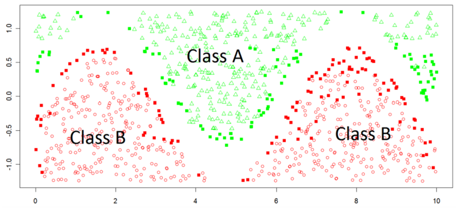
Samia Boukir[1], Wei Feng[2]
sboukir@ipb.fr, wfeng@xidian.edu.cn

[1] Bordeaux INP, G&E Laboratory (EA 4592), F-33600 Pessac, France
[2]Department of Remote Sensing Science and Technology, School of Electronic Engineering, Xidian University, Xian 710071, China

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

OVERVIEW

This work proposes extended bagging algorithms to better handle noisy and multi-class imbalanced classification tasks. These algorithms upgrade the sampling procedure by taking benefit of the confidence in ensemble classification outcome. The underlying idea is that a bagging ensemble learning algorithm can achieve greater performance if it is allowed to choose the data from which it learns. The effectiveness of the proposed methods is demonstrated in performing classification on 10 various data sets.



Low margin instances (filled points) in data set *Sin*

MARGIN

The ensemble margin is an important factor to the generalization performance of voting classifiers. It can be used to measure the degree of confidence of the classification and to guide the design of classification algorithms.

Max-margin

$$margin(x) = \frac{v_y - \max_{c=1,\dots,L \cap c \neq y}(v_c)}{\sum_{c=1}^{L}(v_c)}$$

Sum-margin

$$margin(x) = \frac{v_y - \sum_{c=1,\dots,L \cap c \neq y}(v_c)}{\sum_{c=1}^{L}(v_c)}$$

where v_y is the number of votes for the true class y, v_c is the number of votes for any other class c, and L is the number of classes

- Correctly classified training instances with high margin values represent instances located away from class decision boundaries and can contain a high degree of redundant information. Conversely, training instances with low margin values are often located near class decision boundaries and are more informative in a classification task.
- Misclassified training instances of highest margin (in absolute value) have the highest probability of being mislabelled.

ABOUT BAGGING

The two key ingredients of bagging are *bootstrap* and *aggregation*. Bagging trains a number of base learners, each from a different bootstrap sample, to produce diversity.

We proposed a variant of bagging, ***boundary bagging***, which upgrades the sampling procedure through the ensemble margin (Guo, Boukir and Aussem 2020).

L.Guo, S.Boukir and A.Aussem, "Building bagging on critical instances," Expert Systems, vol. 37, no. 2, p. e12486, 2020.

BOUNDARY BAGGING FOR IDENTIFICATION OF MISLABELLED TRAINING DATA

Algorithm 1 Boundary bagging for noise removal

Inputs:
 Whole training data set S_0 of size N
 Ensemble creation method E
 Base learning algorithm B
 Validation set V
 Percentage M of pruning at each iteration
Initialize $S = S_0$, $i = 0$
 Create an ensemble classifier EB_i with S
 Compute the margin value of each training instance
repeat
 Evaluate EB_i on validation data set V and obtain error rate Er_i of EB_i
 Remove M first highest margin instances that have been misclassified to compose a new cleaner training set S_i
 Set training set S to S_i, $i = i + 1$
 Create an ensemble classifier EB_i with S
until Size of S = 0
Output:
 Best filtered training subset S^* which led to lowest error rate Er^* on V

Data

Data set	Training set	Test set	Variables	Classes
Abalone	2250	1500	8	3
Glass	120	80	10	6
Letter	7500	5000	16	26
Optdigits	1500	1000	64	10
Pendigit	3000	2000	16	10
Segment	1200	800	19	7
Texture	3000	2000	40	11
Vehicle	300	200	18	4
Waveform	3000	2000	21	3
Wine quality-red	900	600	11	6

Data sets from UCI Machine Learning repository

Mislabelled data removal assessment

Data	No filtering	Majority filter	Boundary bagging filter	
			Max-margin	Sum-margin
Abalone	54	54	**54.5**	**54.5**
Glass	**97.5**	**97.5**	97*	96.5
Letter	46.5	48	52.0	**57***
Optdigit	89.5	91	93.5	**94***
Pendigit	90.5	93.0	**95.5**	**95.5**
Segment	92	91	94	**95***
Texture	86.5	89.5	91.5	**94***
Vehicle	72	**73.5**	72.5	73.0*
Waveform	81.5	79.0	**82.5***	82
Wine qua.	**60.5**	60	**60.5**	**60.5**

Classification accuracy of boosting with no filtering, with majority vote and with boundary bagging filtered training sets involving two well-known margins, in presence of 20% of random noise

BOUNDARY BAGGING FOR IMBALANCE SAMPLING

Algorithm 2 Boundary bagging for imbalance sampling

Inputs:
 Whole training data set S of size N
 Number of classes L
 Training data subset S_i of size N_i, $N_1 \leq N_i \leq N_L, \forall$ class i
 Ensemble creation method E
 Base learning algorithm B
 Ensemble size T
 Percentage M_i of pruning $\forall$ class $i > 1$: $M_i = \dfrac{N_i - N_1}{N_i}$
 Sampling rate α
Initialize $t = 1$
 Create an ensemble classifier EB with S
 Compute the margin value of each training instance
repeat
 for $i = 2$ to L
  Remove αM_i first highest margin instances from subset S_i to compose a new subset S_{ti}
  Remove $(1-\alpha)M_i$ instances sampled randomly from subset S_{ti}
 end
 Create a new balanced training set $S_t = S_{t1} \cup \ldots \cup S_{tL}$
 Train a classifier $h_t = B(S_t)$
 Change sampling rate α
until $t = T$
 Fusion of base classifiers $h_t, 1 \leq t \leq T$
 Create ensemble classifier EB_b
Output:
 Ensemble EB_b

Data

Data	Examples	Variables	Classes	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
Cleveland	297	13	5	160	54	35	35	13					
Covtype	8000	54	7	2985	3843	481	33	139	241	278			
Glass	214	10	6	70	76	17	13	9	29				
Hayes-roth	160	4	3	65	64	31							
Newthyroid	215	5	3	150	35	30							
Optdigit	1642	64	10	187	224	196	191	210	197	180	20	197	40
Pendigit	3239	16	10	20	426	408	379	437	397	362	20	394	396
Vehicle	684	17	4	218	50	217	199						
Wilt	4839	5	2	4578	261								
Wine quality-red	1599	11	6	10	53	681	638	199	18				

Imbalanced data sets from UCI Machine Learning repository (*Optdigit*, *Pendigit* and *Vehicle* are artificially imbalanced).

Class imbalance performance

Data	Bagging	Under-Bagging	Boundary bagging	
			Max-margin	Sum-margin
Cleveland	28	29	29	**29.5***
Covtype	32.0	**68**	67.5	**68***
Glass	91.5	93	**93.5***	93
Hayes-roth	77.5	77	80	**83***
Newthyroid	81.5	93.5	94.0	**94.5***
Optdigit	69.5	87.5	**90.5***	90.0
Pendigit	62.5	88.0	**90.5**	**90.5**
Vehicle	71	73	**76.5**	**76.5**
Wilt	87	94.5	**95.5**	**95.5**
Wine qua.	28	**34**	33.5*	33

Average classification accuracy of bagging, UnderBagging and boundary bagging for imbalance sampling involving two well-known margins

Data	Bagging	Under-Bagging	Boundary bagging	
			Max-margin	Sum-margin
Cleveland	100.0	100.0	**92.5***	94.5
Covtype	100.0	**59.0**	68.5	68*
Glass	**20.0**	**20**	**20.0**	**20**
Hayes-roth	52.5	46.5	**31***	32
Newthyroid	38	**12.0**	15.0*	16
Optdigit	100.0	28.5	**20.5**	**20.5**
Pendigit	100.0	29	**27***	29.0
Vehicle	68.5	**56.0**	59	58.5*
Wilt	26.0	7	**4.5**	**4.5**
Wine qua.	100.00	84	**80.5***	83

Maximum classification error per class of bagging, UnderBagging, and boundary bagging for imbalance sampling involving two well-known margins

CONCLUSION

Results from this study show that our extended bagging approach for mislabelled training data filtering outperforms the majority vote noise filter.

Our experiments also demonstrate the superiority of our extended bagging approach in handling the class imbalance learning problem compared with traditional bagging and UnderBagging.