



**HAL**  
open science

## Data annotation with active learning: application to environmental surveys

Chloé Friguét, Romain Dambreville, Ewa Kijak, Mathieu Laroze, Sébastien Lefèvre

### ► To cite this version:

Chloé Friguét, Romain Dambreville, Ewa Kijak, Mathieu Laroze, Sébastien Lefèvre. Data annotation with active learning: application to environmental surveys. JDS 2020 - 52èmes Journées de Statistique de la Société Française de Statistique (SFdS), May 2020, Nice, France. pp.1-6. hal-03445740

**HAL Id: hal-03445740**

**<https://hal.science/hal-03445740>**

Submitted on 24 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# DATA ANNOTATION WITH ACTIVE LEARNING: APPLICATION TO ENVIRONMENTAL SURVEYS

Chloé Friguét<sup>1</sup>, Romain Dambreville<sup>2</sup>, Ewa Kijak<sup>3</sup>, Mathieu Laroze<sup>4</sup> & Sébastien Lefèvre<sup>5</sup>

<sup>1</sup> *Univ. Bretagne-Sud, UMR 6074, IRISA, Vannes, France - chloe.friguet@irisa.fr*

<sup>2</sup> *WIPSEA, Rennes, France - romain.dambreville@wipsea.fr*

<sup>3</sup> *Univ. Rennes, UMR 6074, IRISA, Rennes, France - ewa.kijak@irisa.fr*

<sup>4</sup> *Univ. Rennes, UMR 6074, IRISA et WIPSEA, Rennes, France - mathieu.laroze@irisa.fr*

<sup>5</sup> *Univ. Bretagne-Sud, UMR 6074, IRISA, Vannes, France - sebastien.lefevre@irisa.fr*

**Résumé.** Une procédure d'apprentissage actif est proposée pour réduire le coût d'annotation d'images aériennes pour des suivis environnementaux. La sélection des instances à étiqueter à chaque étape du processus actif est contrainte à l'appartenance à un groupe, une image (ou une partie d'image) dans notre cas. Un score pour classer les images et identifier celle qui doit être annotée à chaque itération est défini, en fonction de l'incertitude et des performances de détection du classifieur. Les performances de plusieurs stratégies concernant le gain d'interaction avec l'utilisateur sont discutées à partir d'une expérience sur des données d'images réelles collectées pour une étude environnementale.

**Mots-clés.** Apprentissage actif, annotation d'images aériennes, suivis environnementaux

**Abstract.** An active learning framework is introduced to deal with reducing the annotation cost for aerial images in environmental surveys. The selection of the queried instances at each step of the active process is here constrained by requiring that they belong to a group, an image (or a part of it) in our case. A score to rank the images and identify the one that should be annotated at each iteration is defined, based on both classifier uncertainty and performances. The performances of several strategies regarding the interaction gain are discussed based on an experiment on real image data collected for an environmental survey.

**Keywords.** Active learning, aerial images labelling, environmental surveys

## 1 Active learning and selective sampling

Machine Learning (ML) aims at deriving algorithms that can automatically learn from available data and make predictions. Active Learning (AL) is related to semi-supervised ML in which a learning algorithm is able to interact with the user to get some information about the label of new data during the training step. It is motivated by situations in which it is easy to collect unlabelled data but costly (time, money, tedious task) to (manually) obtain their labels. It stems from the idea that we should only acquire labels that actually improve our ability to make accurate predictions. More formally, a supervised model  $\Theta$

is trained incrementally on a training dataset  $X$ . A query criterion  $Q$  searches over the unlabelled dataset  $\mathcal{U}$  and queries an oracle  $O$  to get a label feedback for a selected instance  $\mathbf{x}^*$ . The new labelled instance is added to the labelled dataset  $\mathcal{L}$  and removed from  $\mathcal{U}$ . The model  $\Theta$  is then re-trained on the augmented labelled dataset and the process is repeated until an ending criteria is met. Instances that are more useful than others for some performances have to be identified to create an optimal training dataset: well chosen, fewer representative instances are needed to achieve a similar performance. This selection process has been investigated as *selective sampling* [5]. The importance of an instance is related to a high level of both the information and uncertainty relatively to the trained model, considering therefore a trade-off between *informativeness* (ability to reduce the uncertainty of a statistical model) and *representativeness* (ability to represent the whole input data space) of the selection process [1].

## 2 Active learning to ease annotation

**Motivating context: labelling objects in aerial images** Nowadays, remote sensing technologies greatly ease environmental assessment over large study areas using aerial images, e.g. for monitoring and counting animals or ships. In the fields of both machine learning and image processing, many algorithms have been developed to tackle the complex task of object detection and to fasten and automate the counting processes. In practice, each image is divided into patches and object detection can be restated as a binary classification issue, to predict if the object of interest is on a patch or not, over the whole set of images  $\mathcal{S}$ . Most of these procedures are then supervised, and need to have prior ground truth available for each patch. However, manually labelling the patches requires, even for an expert, a time-consuming and tedious process. To assist the annotation process, an active learning approach can be used [6, 4], allowing interaction with the expert such as label confirmation or correction for each patch, at the query step. Note that in our context of environmental surveys, the data are unbalanced: there are only a few objects on each image, so most of patches are negatives. Moreover, the patch labelling is not easy even for an expert, and this query for a single patch usually requires some contextual information through the visualisation of the surrounding patches, that can therefore also be labelled at the same step. To address this challenge, the proposed method aims at assisting the annotation process by introducing an active learning procedure, querying the expert with groups of patches taken from the same image to ease annotation.

**Proposed method** The proposed active process is detailed with pseudo-code in Algo. 1. In our context, the initial data are a set  $\mathcal{S} = \{I_0; \dots; I_K\}$  of  $K + 1$  images, each one being composed  $N$  patches mapped into a set of  $p$  features (see Figure 1). The input of the active selection algorithm is then a  $(K + 1)$ -set of  $N \times p$ -matrix of instances  $x_i^k \in \mathbb{R}^p$ . The output is a class value for each instance denoted  $y_i \in \mathcal{Y} = \{0; 1\}$ , 1 being the positive class.

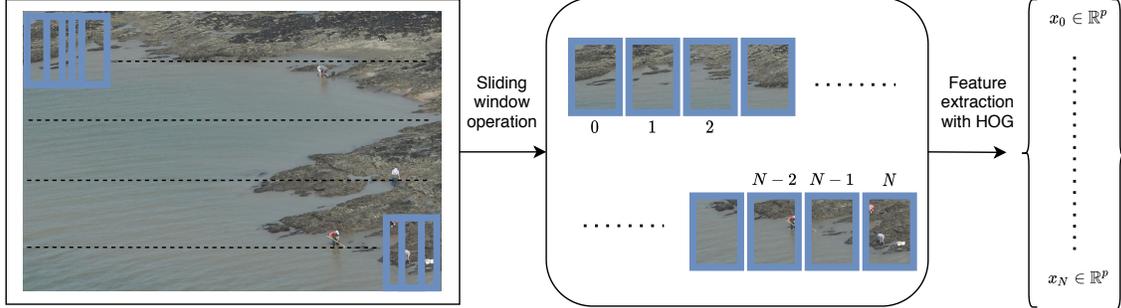


Figure 1: From image to patches and features extraction. The  $N$  patches from image  $k$  are considered together in the annotation process

Let us denote by  $\mathcal{U}_\ell$  (resp  $\mathcal{L}_\ell$ ) the set of unlabelled (resp. labelled) instances at iteration  $\ell$ . At the beginning,  $\mathcal{L}_0 = \{x_i^0; y_i^0\}_{i=1\dots N}$  denotes the labelled data for initialisation and  $\mathcal{U}_0$  is the set of remaining data to be labelled.

---

**Algorithm 1:** Main steps of the proposed active learning algorithm

---

- Input** : Initial training set  $\mathcal{L}^0$   
**Input** : Pool of unlabelled candidates  $\mathcal{U}^0$
- 1 Initiate  $\ell = 0$
  - 2 **repeat**
  - 3     Train a classifier  $\mathcal{C}$  with current training set  $\mathcal{L}^\ell$
  - 4     Predict labels  $y$  for all instances in current unlabelled set  $\mathcal{U}^\ell$  with  $\mathcal{C}$
  - 5     **for each image  $k$  candidate in  $\mathcal{U}^\ell$  do**
  - 6         Compute the score  $s_k$  according to (1) and predictions by  $\mathcal{C}$
  - 7     Rank candidates in  $\mathcal{U}^\ell$  according to the scores  $s_k$
  - 8     Select the most interesting image  $I^* = \operatorname{argmax}_k \{s_k\}$
  - 9     Query  $I^*$  to the oracle  $O$  to receive labels for all its instances: the oracle confirms positive labels and/or correct false negative ones.
  - 10    Add labelled instances to the training set:  $\mathcal{L}^{\ell+1} = \mathcal{L}^\ell \cup I^*$  and  $\mathcal{U}^{\ell+1} = \mathcal{U}^\ell \setminus I^*$ .
  - 11     $\ell = \ell + 1$
  - 12 **until**  $\mathcal{S}$  has been fully annotated;
- 

In usual algorithms, at each iteration, queried instances are selected one after another on criteria based on their informativeness or their uncertainty. In our context, the observed instances to be labelled are grouped into consistent sets (images) that we propose to process as a whole in the active query. This strategy necessitates therefore to define a global score to rate the relevance of each image to be the next one be queried to get all the labels of its patches. We propose a selective sampling query taking into consideration two main criteria: (i) Certain Instances predicted positive (resp. negative) by the classifier with a probability greater than a fixed threshold  $t_c$  (resp. lower than  $1 - t_c$ ); (ii) Uncertain Instances predicted either positive or negative by the classifier with a probability lying in

$[0.5 \pm t_{unc}]$ , for a fixed threshold  $t_{unc}$ . The global score is defined as the harmonic mean combining the number of positive certain instances  $PCI$  and all uncertain instances  $UI$ :

$$s = \frac{PCI \times UI}{PCI + UI} \quad (1)$$

The data are strongly imbalanced in our context: consequently, at each step, there are much more positive instances than negatives ones that are labelled. Learning from such data requires appropriate training strategies and metrics to assess the algorithm performances [2]. Besides, selecting the effective training subset at each step is not trivial and must be done very carefully to keep representative instances in the majority class. In the following, we consider several under-sampling strategies, based both on user interactions and classifier confidence: (UC) a balanced subset of uncertain instances; (UC+C) a balanced subset of uncertain and certain instances; (UC+C+EK) a balanced subset of uncertain and certain instances enriched with Extra-Knowledge (all instances corrected by the oracle that would not have been selected with the previous strategies: false positives, false negatives and positives with medium confidence).

In the imbalanced learning case, usual performance evaluation metrics are based on positives detection (precision, recall and F-score). In our context, we also consider a user-oriented criterion, based on the number of interactions. Interactions with the expert for the label annotation process can be either to correct the false positives or to add the missing positives (i.e. correct the false negatives). True positives and true negatives are validated implicitly. The evolution of the total number of interactions over the iterations of the algorithm is considered in the following as an evaluation criterion of the method: the method succeeds if, at the end of the algorithm, it is less than the number of interactions needed to annotate the full dataset (total number of true positives  $N^+$ ). The gain in interaction  $G_K$  through our method is calculated by their difference.

### 3 Experiments

**Settings** The experiments are carried out on a set of aerial images of humans gathering shellfish on the seashore in the Natural Park of Morbihan (South Brittany, France) during spring tides [3]. The aim is to evaluate the number of people on the seashore in this period of high attendance and deduce the pressure of their activity on the environment. Table 1 reports the dataset characteristics and 3 images are shown in Figure 2. Instances (patches) are extracted with a sliding window of size  $64 \times 32$  with a stride of 8 and Histogram of Oriented Gradients (HOG) features are extracted. The classification task at each iteration is performed considering Support Vector Machine (SVM). A 4-fold training-test setting is considered to assess the performance of the algorithm. The procedure may depend upon the image chosen for initialisation, hence averages computed over all possible initialisations are reported, for each criterion. Thresholds for certain and uncertain data are empirically set to  $t_c = 0.8$  and  $t_{unc} = 0.1$ .



Figure 2: Images from shellfish dataset with 15, 3 and 26 shellfish gatherers

Table 1: Dataset statistics (average over the 4 train/test splits)

	Training set	Test set
# images	23	5
# total instances	682,410	148,350
# positive instances	537 ( $N^+$ )	114

**Main results** The proposed method is first classically evaluated regarding the classifier performances *w.r.t.* the retraining strategies (see Table. 2). Adding diversity to the uncertain training (UC) set considering also certain examples (UC+C) allows better detection performances (higher F-score) but taking into account the extra-knowledge (UC+C+EK) decreases the average F-score on the test sets: if the precision increases, the recall sharply decreases. The number of interactions with the user is the number of corrections that have to be made by the user to the classifier predictions (FP, FN) during successive iterations  $\ell = 1..K$ . The interaction gain  $G_\ell$  is the difference with the number of true positives, that would have been the number of interactions in an annotation task without any selective sample process. A crucial characteristic for an active process is therefore the total number of interaction gain at the end of the process ( $G_K$ ). According to this criterion, UC+C+EK performs best as it is more robust to the initialisation step. This conversely differs from classification performances, but this user metric evaluation fits the objective of easing the annotation.

Table 2: Evaluation of re-training strategies: classifier performances and user interaction gain - Average (std) over the 4 test sets.

	Classifier perf.			Interaction gain ( $G_K$ )	
	F-score	Recall	Precision	mean (std)	min/max
UC	0.22 (0.17)	0.58 (0.30)	0.25 (0.25)	20.5 (26.5)	0/87
UC+C	<b>0.42</b> (0.10)	<b>0.76</b> (0.07)	0.36 (0.13)	41.8 (38.9)	0/ <b>135</b>
UC+C+EK	0.08 (0.04)	0.05 (0.03)	<b>0.50</b> (0.20)	<b>77.5</b> (16.6)	<b>43</b> /102

## 4 Conclusion and discussion

We introduced an active learning annotation process to reduce the annotation cost when creating a ground truth. Usual active learning algorithms perform instances selection from the whole set of input data. In the present work, the selection of the queried instances is constrained by requiring that they belong to a group, i.e. (a part of) an image here, to ease the annotator task as the queried instances are proposed in their comprehensive context. We defined a score to rank the images and identify the one that should be annotated at each iteration, based on both uncertainty and true positives. The main objective is to reduce the number of human interactions on the overall process, starting from a first annotated image, rather than reaching the maximum final accuracy. Therefore, the annotation cost is measured through the gain in interactions (corrections of the classifier decisions by the annotator) with respect to a labelling task from scratch. At each iteration, the classifier is retrained according to a specific subset of data. Several strategies have been compared and their performances regarding the interaction gain have been discussed. We also highlight that initialisation is a crucial step to our design. While out of the scope of this study, it requires further investigation to gain robustness in the process. Improvements can also be brought considering more appropriate features, such as those based on convolutional auto-encoder that can suit better than HOG for small object detection problem in aerial images.

## References

- [1] Y. Fu, X. Zhu, and B. Li. A survey on instance selection for active learning. *Knowledge and Information Systems*, 35(2):249–283, 2013.
- [2] H He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.
- [3] M. Laroze, L. Courtrai, and S. Lefèvre. Human detection from aerial imagery for automatic counting of shellfish gatherers. In *Inter. Conf. on Computer Vision Theory and Applications (VISAPP)*, 2016.
- [4] M. Laroze, R. Dambreville, C. Friguet, E. Kijak, and S. Lefèvre. Active Learning to Assist Annotation of Aerial Images in Environmental Surveys. In *International Conference on Content-Based Multimedia Indexing*, pages 1–6, 2018.
- [5] B. Settles. Active learning literature survey. Computer sciences technical report, University of Wisconsin–Madison, 2010.
- [6] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011.