Datasets with Rich Labels for Machine Learning

Arthur Hoarau, Constance Thierry, Arnaud Martin, Jean-Christophe Dubois, Yolande Le Gall

Datasets with rich labels for machine learning

Arthur Hoarau
Univ Rennes, CNRS, IRISA, DRUID
Rennes, France
arthur.hoarau@univ-rennes.fr

Constance Thierry
Univ Rennes, CNRS, IRISA, DRUID
Rennes, France
constance.thierry@univ-rennes.fr

Arnaud Martin
Univ Rennes, CNRS, IRISA, DRUID
Rennes, France
arnaud.martin@univ-rennes.fr

Jean-Christophe Dubois
Univ Rennes, CNRS, IRISA, DRUID
Rennes, France
jean-christophe.dubois@univ-rennes.fr

Yolande Le Gall
Univ Rennes, CNRS, IRISA, DRUID
Rennes, France
yolande.le-gall@univ-rennes.fr

***Abstract*—Most datasets used for classification use hard labels. In this paper, five new datasets labeled with uncertainty and imprecision by crowdsourcing contributors are presented. Richer labels are modeled with the theory of belief functions, which generalizes several reasoning frameworks with uncertainty, such as possibilities or probabilities. These datasets can be used with classical models using hard labels but also with probabilistic, fuzzy or even evidential models. Several concrete application cases are presented, for which these new datasets provide a useful knowledge representation of the user's uncertainty.**

***Index Terms*—Crowdsourcing, Datasets, Evidential Learning, Belief Functions**

I. Introduction

Crowdsourcing [1] is the outsourcing of a task to a crowd of contributors on dedicated platforms. Several types of platforms exist [2] depending on the proposed type of task, the expertise of the crowd that performs them and the remuneration granted. For example, on routine activity platforms, the tasks are simple, do not require expertise and the crowd is very diversified. Crowdsourcing is as well used by companies to easily and quickly call on manpower [3], as by research organizations, especially for data annotation [4].

An interface for crowdsourcing was introduced in [5] allowing users to label data in an uncertain and imprecise way. This paper proposes to use this interface to obtain datasets with rich labels, this time for the machine learning community. The theory of belief functions is used here as in [5] but the combination of information is different. This framework allows to generalize probabilities but also other theories for reasoning with uncertainty, like imprecise probabilities or possibilities.

With these new datasets, the goal is to provide the ability for uncertainty-based models to access data that have actually been labeled by contributors, in an uncertain and imprecise manner. The paper is organized as follows, section II reviews richer labels and belief function theory. Section III introduces the five new datasets and section IV illustrates areas of application where such datasets, because of their novelty, could provide solutions. Finally Section V concludes the article.

II. Rich labels and theory of belief functions

A. Rich labels

Most of the datasets used for classification consider hard labels, with a binary membership where the observation is either a member of the class or not. In this paper, we refer as rich labels the elements of response given by a source that may include several degrees of imprecision. Philippe Smets hypothesized that the more imprecise humans are, the more certain they are [6], [7]. To our knowledge, there is no dataset for machine learning labeled in an uncertain and imprecise way by contributors who have had the opportunity to represent this imperfection. Most of the time, datasets with hard labels are noisy or, as in [8], fuzzy labels are extracted from the original dataset. Other datasets directly reference fuzzy labels [9], but the imprecision is related to the observations and none of them represent several degrees of user imprecision. Uncertain and imprecise answers are given in [5] but these data are not suitable for machine learning, with several contributors labeling the same picture, and especially the number of observations is much too small. In this document, these labels are called rich - but could also be named uncertain, soft or imperfect depending on the context - as opposed to hard labels and they are modeled using the theory of belief functions.

B. Theory of belief functions

The theory of belief functions, also called Dempster-Shafer theory [10], [11], is used in this study to model uncertainty and imprecision in the labels.

Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the frame of discernment for M exclusive and exhaustive hypotheses. The power set 2^Ω is the set of all subsets of Ω. A Basic Belief Assignment is the belief that a source may have about the elements of the power set of Ω, this function assigns a mass to each element of this power set such that the sum of all masses is equal to 1.

$$\begin{aligned} &m : 2^\Omega \rightarrow [0,1], \\ &\sum_{A \in 2^\Omega} m(A) = 1. \end{aligned} \tag{1}$$

Each subset $A \in 2^\Omega$ such that $m(A) > 0$ is called a *focal element* of m. The uncertainty is therefore represented by a mass $m(A) < 1$ on a focal element A and the imprecision is represented by a non-null mass $m(A) > 0$ on a focal element A such that $|A| > 1$.

A mass function m is called *simple support mass function* when it has two focal elements, one of which is Ω:

$$\begin{aligned} m(A) &= 1 - w,\ A \in 2^\Omega, \\ m(\Omega) &= w, \\ m(B) &= 0,\ B \in 2^\Omega \backslash \{A, \Omega\}. \end{aligned} \tag{2}$$

with $w \in [0, 1]$, the mass function m can then be noted A^w.

The cautious rule of combination $\textcircled{\wedge}$, introduced by Denœux in [12], allows to combine the information contained in two dependent mass functions. It is defined as follows:

$$m^1 \textcircled{\wedge} m^2 = \bigoplus_{A \in 2^\Omega} A^{w_1(A) \wedge w_2(A)} \tag{3}$$

with $\wedge$ the operator of minimum and the sign $\bigoplus$ is Dempster's rule of combination [11]. This combination rule will be used to model the labels.

III. Credal datasets

A. Imprecise and uncertain contributors

In this paper, we propose five new datasets. The adopted approach is first explained and then each dataset is described. With the interface introduced in [7], crowdsourcing contributors are allowed to label observations in an uncertain and imprecise way (see [6] for definitions). The user expresses uncertainty on a Lickert scale (from 1 to 7 levels). Its imprecision is expressed by a selection of multiple propositions. Each experiment is performed on a crowd of non-specific and paid contributors[1]. If a user is not completely precise and certain, a second step allows him; *-to refine* his answer when he has selected more than one class, or *-to choose more classes* when he has chosen a single answer with a non maximum certainty. The two answers are modeled within the framework of belief functions and combined using the cautious combination rule.

Figure 1 presents the interface proposed to the contributor to label each of the observations. Several proposals can be selected by the user who must also estimate the global certainty of his answer.

TABLE I
Credal Datasets

Name	Classes	Pictures	Contributors	Features
Credal Dog-7	7	700	50	43
Credal Dog-4	4	400	50	47
Credal Dog-2	2	200	50	42
Credal Bird-10	10	200	50	30
Credal Bird-2	2	40	50	17

Table I shows the five new proposed datasets. The content made available for each of them is presented as follows:

- **Features:** *Pictures* in RGB[2]. *Large feature vector* of 512 variables. *Feature vector* of fewer components of a PCA[3].
- **Classes:** *True class* representing the ground truth. *Rich labels* from the contributors' responses.
- **Raw crowdsourcing inputs:** to bypass belief functions and get direct user responses on the interface.

The process of labeling is highlighted with a few examples in the following sections, describing each dataset.

B. Credal Dog-7

The first dataset represents pictures of 7 breeds of dogs. Exactly 512 features (also available in this dataset) are extracted from each observations[4] (*i.e.* the 400x400 RGB pictures of dogs) and a PCA is carried out. The first components regrouping in total 70% of the variance are retained, these are the 43 selected features. Note that for all datasets only numerical variables are used, the dataset is described as follows. *Observations:* 700, *Classes:* 7, *Features:* 43, *Contributors:* 50, *Labels by contributors:* 14.

An example of the process of labeling an observation in this dataset is shown in Figure 1. Two steps are presented, the first step asks the user to choose as much breeds as possible in order to have a high certainty. Here, *Brittany*, *Shetland Sheepdog* and *Beagle* are selected with a certainty of 6 out of 7. The second step allows the user to reduce his choice, by doing so, his certainty is bound to decrease. In the opposite case, where the user first chooses a precise answer, the second step asks the user not to reduce his choice, but to enlarge if possible the number of selected classes. If in the first step the user has a precise (only one class) and certain (totally certain) answer, no second step is offered. To obtain the mass associated with the response, the selected certainty value is divided by the maximum value[5], this mass is assigned to the focal element of the selected classes and the rest goes to ignorance. Two mass functions result from the user's first and second choice giving m_1 and m_2 respectively:

- m_1**:** $m_1(\{\omega_1, \omega_2, \omega_3\}) = 0.86$, $m_1(\Omega) = 0.14$
- m_2**:** $m_2(\{\omega_1\}) = 0.43$, $m_2(\Omega) = 0.57$

with $\omega_1 = Brittany$, $\omega_2 = Shetland\ Sheepdog$ and $\omega_3 = Beagle$. The following label m given to the observation is the cautious combination of the two masses:

- m**:** $m(\{\omega_1\}) = 0.43$, $m(\{\omega_1, \omega_2, \omega_3\}) = 0.49$, $m(\Omega) = 0.08$

The labels in these datasets are therefore given in this form and each observation is assigned a mass function by a user.

The 2-dimension representation in Figure 2 is the dataset on the first two principal components of a PCA. It is clear that metaclasses are formed, for example *Shetland Sheepdog* and *Collie* are very close. Indeed, these two breeds of dogs are

[1]The experiments were carried out in France, hence the presence of dog breeds and bird species likely to be known to Europeans.

[2]On recent systems color values of pixels are encoded on Red, Green, and Blue components.

[3]Principal Component Analysis, in data analysis.

[4]A ResNet-34 is trained and during the prediction phase, for each observation, the 512 outputs of the last layer give the features.

[5]On the Likert scale used, the values range from 1 to 7.

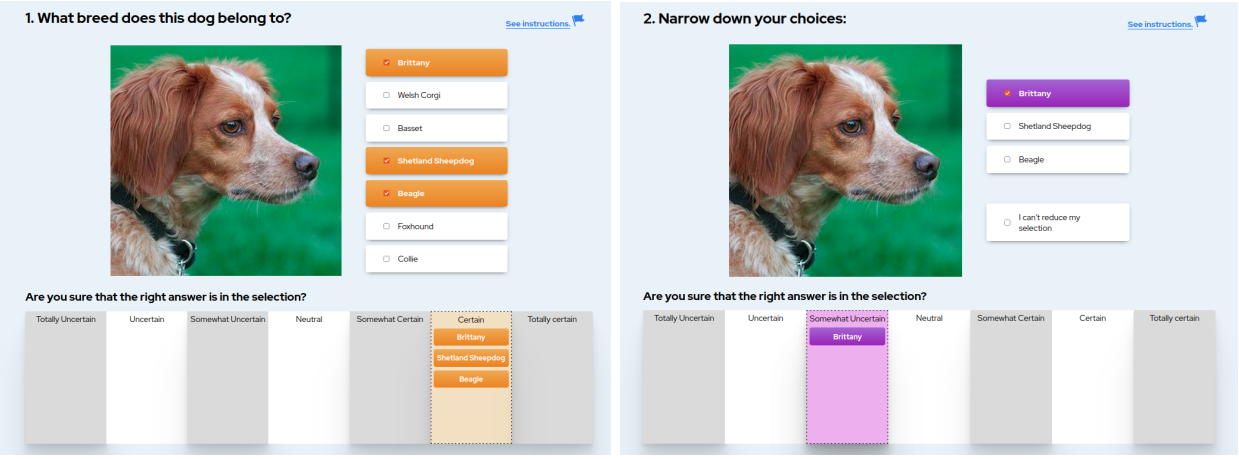

Fig. 1. Labeling process for a crowdsourcing contributor on the Credal Dog-7 dataset.

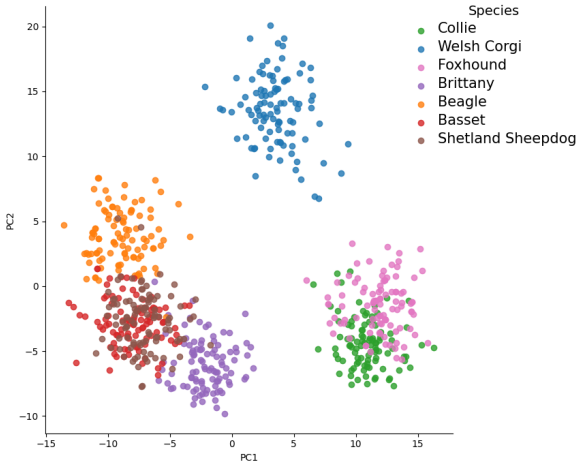

Fig. 2. Principal component analysis on Credal Dog-7, with a representation of the dataset on the first principal component (PC1) and the second (PC2).

very similar, hardly discernible to an inexperienced user. The *Welsh Corgi* is left a bit apart, while the remaining 4 breeds form a cluster.

Half of the collected responses used 2 steps (the other half corresponds to the answers where the user announces that he knows perfectly the breed, or that he can't answer the iteration). Also, 80% of the *totally certain* responses contain the true class, and 79% of both certain and precise answers are correct. This information is important, it allows us to see that the users who say they know the real class often have the right answer, they are the experienced users.

C. Credal Dog-4

This dataset has great similarities with the first one, but is a different dataset, both the extracted features and the labels are different, from different contributors. Similarly, 512 features are extracted and a PCA yields 47 principal components retaining 70% of the total variance. *Observations:* 400, *Classes:* 4, *Features:* 47, *Contributors:* 50, *Labels by contributors:* 8.

The interface for the Credal Dog-4 labeling campaign is showed in Figure 3. The same two steps are presented, and here, *Foxhound* and *Beagle* are selected with a certainty of 7 out of 7, which means that the user knows that the dog is one of these two breeds. In the second step, the user chooses *Foxhound* with a high certainty of 6 out of 7, which means that among his previous selection, he has a strong belief that the true class is *Foxhound*. The two mass functions m_1 and m_2 resulting from the answers are given as follows:

- m_1: $m_1(\{\omega_1, \omega_2\}) = 1$
- m_2: $m_2(\{\omega_1\}) = 0.86$, $m_2(\Omega) = 0.14$

with $\omega_1 = Foxhound$ and $\omega_2 = Beagle$. The following label m given to the observation is the cautious combination of the two masses:

- m: $m(\{\omega_1\}) = 0.86$, $m(\{\omega_1, \omega_2\}) = 0.14$

Figure 4 is the representation of the dataset on the PCA's first two principal components. The two breeds *Foxhound* and *Beagle* are very close, and are also often difficult to differentiate. While *Brittany* and *Basset* seem to belong to their own clusters.

For each experience, the results seem to agree, there are always about half of the responses (45%) that are two-step and about 20% of wrong answers when users specify a certain and precise answer. Here, 81% of the certain answers are correct, and 78% of the certain and precise answers are correct.

D. Credal Dog-2

The last 2-class dataset on dog breeds obtained during crowdsourcing is defined as follows. *Observations:* 200, *Classes:* 2, *Features:* 42, *Contributors:* 50, *Labels by contributor:* 4. Having two classes and only one possible degree of imprecision is not representative of what can be done with richer labels. But often in machine learning, simple 2-class datasets are used to introduce new models or definitions. Credal Dog-2 is then proposed to fill this need. Figure 5 shows the process of labeling for an observation in this dataset. Only

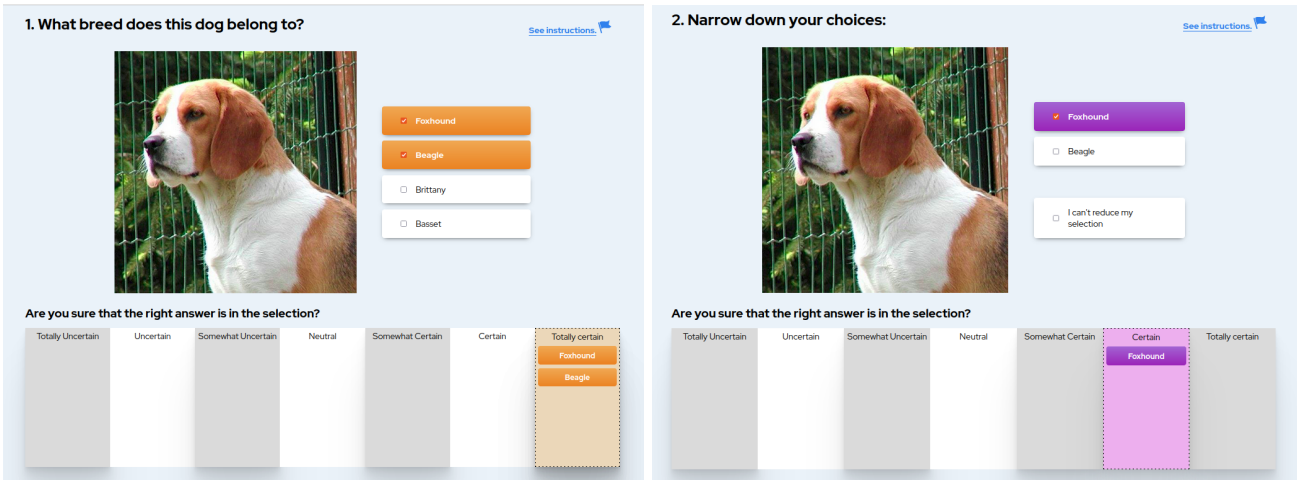

Fig. 3. Labeling process for a crowdsourcing contributor on the Credal Dog-4 dataset.

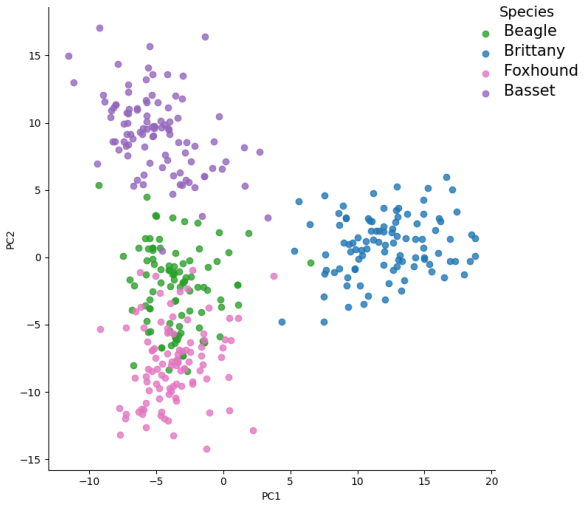

Fig. 4. Principal component analysis on Credal Dog-4, with a representation of the dataset on the first principal component (PC1) and the second (PC2).

two breeds are possible, and for the first step, both *Brittany* and *Beagle* are selected with a certainty of 5 out of 7[6]. In the second step, the user chooses *Beagle* with a lower certainty of 2 out of 7, which means that among his previous selection, he has not a strong belief that the true class is *Beagle* but is still more confident than for *Brittany*. The two mass functions m_1 and m_2 resulting from the answers are given as follows:

- m_1: $m_1(\Omega) = m_1(\{\omega_1, \omega_2\}) = 1$
- m_2: $m_2(\{\omega_1\}) = 0.29$, $m_2(\Omega) = 0.71$

with $\omega_1 = Beagle$ and $\omega_2 = Brittany$. The following label m given to the observation is the cautious combination of the two masses:

- m: $m(\{\omega_1\}) = 0.29$, $m(\{\omega_1, \omega_2\}) = 0.71$

E. Credal Bird-10

This dataset is obtained from the same interface but this time the experience is done over ten species of birds. *Observations:* 200, *Classes:* 10, *Features:* 30, *Contributors:* 50, *Labels by contributor:* 20.

Another specificity of this dataset is that *Labels by contributor* $\times$ *Contributors* $\neq$ *Observations*. In the previous datasets, each observation is labeled once, and each contributor label a fixed number of observations. During this experience, multiple users labeled the same observation, and one of the labels is randomly selected to be the final label in order to have one label per picture. The rest is the same, 512 features are extracted and a PCA yields 30 principal components retaining 70% of the total variance.

The interface for the campaign is showed in Figure 7. First *Western Jackdaw*, *Carrion Crow*, *Common Raven* and *Rook* are selected with a certainty of 7 out of 7. In the second step, the user chooses *Western Jackdaw* and *Common Raven* with a certainty of 4 out of 7. The two mass functions m_1 and m_2 resulting from the answers are given as follows:

- m_1: $m_1(\{\omega_1, \omega_2, \omega_3, \omega_4\}) = 1$
- m_2: $m_2(\{\omega_1, \omega_2\}) = 0.57$, $m_2(\Omega) = 0.43$

with $\omega_1 = Western\ Jackdaw$, $\omega_2 = Common\ Raven$, $\omega_3 = Carrion\ Crow$ and $\omega_4 = Rook$. The following label m given to the observation is the cautious combination of the two masses:

- m: $m(\{\omega_1, \omega_2\}) = 0.57$, $m(\{\omega_1, \omega_2, \omega_3, \omega_4\}) = 0.43$

On this label, there is two degrees of ignorance, one on $\{\omega_1, \omega_2\}$ and one on $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ but there is no uncertainty on the total ignorance because $m(\{\Omega\}) = 0$.

The 2-dimension representation in Figure 8 also shows clear metaclasses. Species *Marsh Tit*, *Great Tit* and *Coal Tit* are put together. It is the same for *Common Raven*, *Western Jackdaw*, *Rook* and *Carrion Crow*. Only *European Robin* remains alone.

[6]In this experiment, the first-step certainty does not matter if the user selects both classes, as it represents total ignorance.

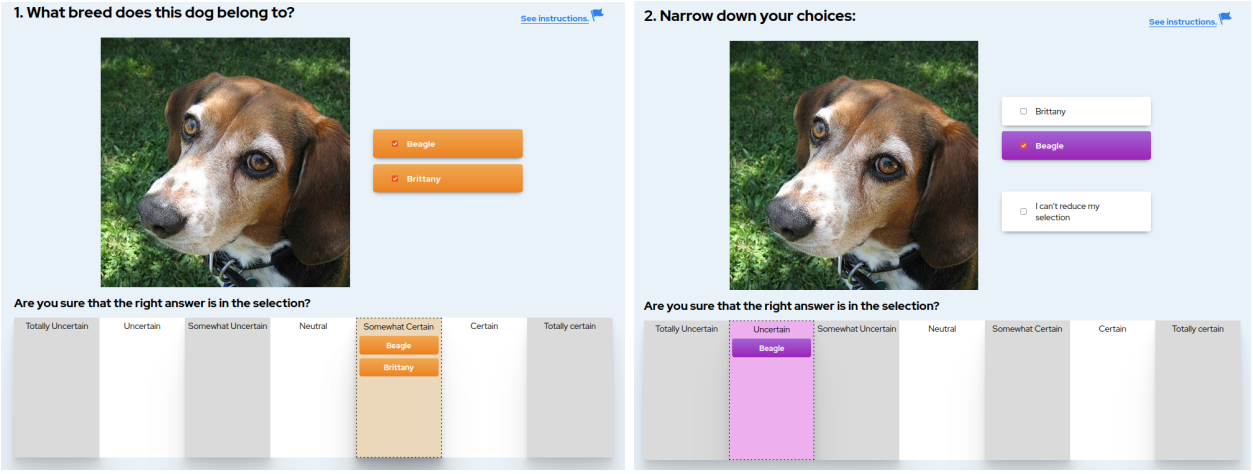

Fig. 5. Labeling process for a crowdsourcing contributor on the Credal Dog-2 dataset.

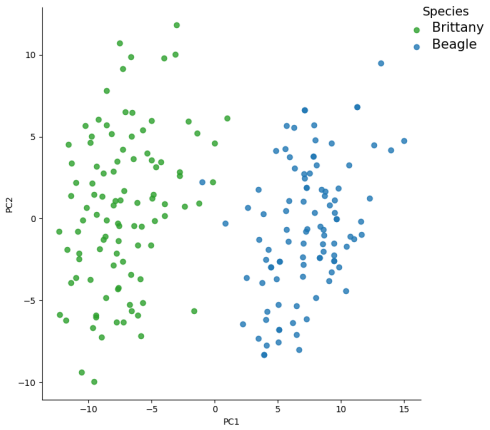

Fig. 6. Principal component analysis on Credal Dog-2, with a representation of the dataset on the first principal component (PC1) and the second (PC2).

For the Credal Bird-10 dataset, the proportions of iterations (46%) and good answers (82% for max certainty) are almost the same, with more correct *certain and precise* answers (91%).

F. Credal Bird-2

The last dataset is a small one, also with bird species and on only two classes. *Observations:* 40, *Classes:* 2, *Features:* 17, *Contributors:* 50, *Labels by contributor:* 4. Specificities for this dataset are the same as in the previous one, and is based on the same combination than for Credal Dog-2.

IV. Applications

This section presents several applications for this kind of dataset. From supervised and unsupervised learning to active learning, they can be used in a large spectrum of applications and within many frameworks.

A. Evidential learning

Thanks to the use of the theory of belief functions; probabilities, possibilities, credal sets and fuzzy sets can be derived from mass functions. These datasets are compliant with evidential classifiers [13]–[15] but also with a large number of models such as fuzzy or imprecise classifiers.

B. Increasing performance

The idea of getting closer to what the user really thinks may increase performance. Indeed, it would be better to know that someone is unsure of its wrong answer than just having a wrong *hard* label. Those kinds of datasets can help to show that by giving the possibility to users to answer in an uncertain and imprecise way, they can give more reliable information than with *hard* labels. In the experiment in Table II, each crowdsourcing campaign is reiterated but this time a single *hard* label is required from contributors. These *hard* labels are also present in the available resources from each dataset. The Evidential K-Nearest Neighbors [14] is used for classification, with a 5-fold cross validation to estimate the best K. Mean accuracies over 100 experiments are presented as performance score. This experiment shows that, overall, giving the possibility to users to answer imperfectly can increase performance.

TABLE II
Mean accuracy by dataset for hard and rich labels ($\pm$ a 95% confidence interval for the estimation of the mean).

Datasets	Hard labels	Rich labels
Credal Dog-7	68.7 $\pm$ 0.8	75.8 $\pm$ 0.7
Credal Dog-4	70.8 $\pm$ 1.0	69.3 $\pm$ 1.0
Credal Dog-2	98.4 $\pm$ 0.5	98.0 $\pm$ 0.4
Credal Bird-10	52.8 $\pm$ 1.5	60.7 $\pm$ 1.5
Credal Bird-2	51.6 $\pm$ 3.1	63.5 $\pm$ 3.7

C. Knowledge representation

The knowledge added in the labels can be modeled and the representation of uncertainty can be useful, for example in active learning [16]. A sampling by uncertainty then allows to find the zones of uncertainty in the space of the features.

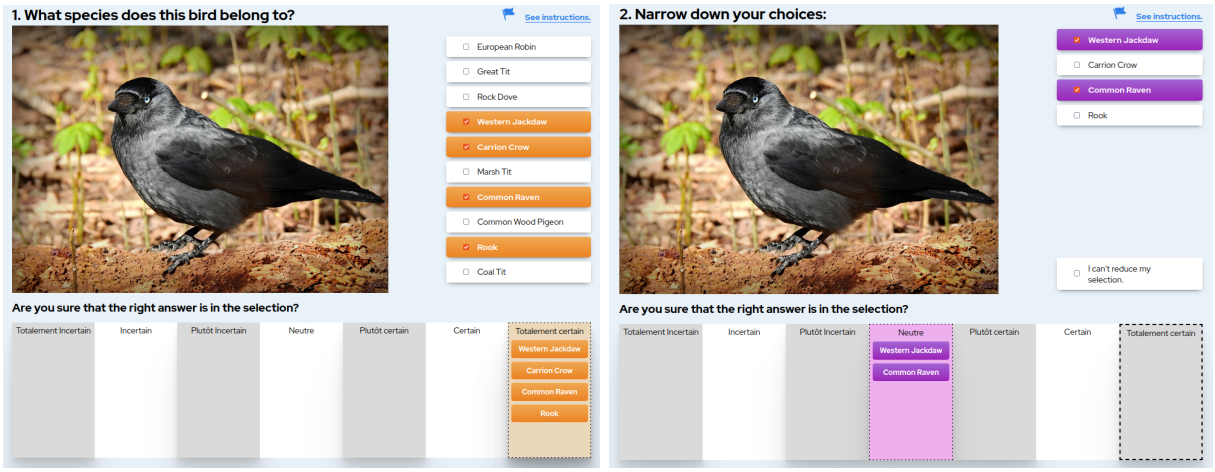

Fig. 7. Labeling process for a crowdsourcing contributor on the Credal Bird-10 dataset.

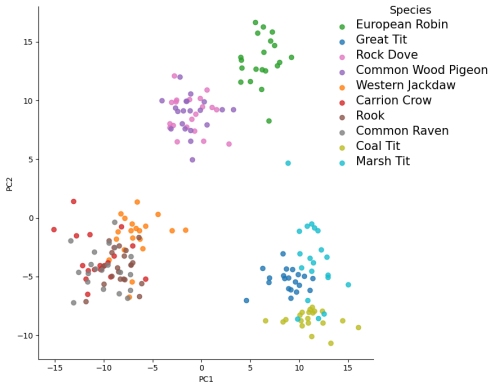

Fig. 8. Principal component analysis on Credal Bird-10, with a representation of the dataset on the first principal component (PC1) and the second (PC2).

V. Conclusion

In this paper, we proposed five new datasets, labeled in an uncertain and imprecise way by crowdsourcing contributors. During the data collection, two steps are proposed to the contributor to label an image: he first formulates an initial answer, which he can then refine or enlarge according to the precision but also the certainty given to his proposal. The information collected is modeled by the theory of belief functions. It provides flexibility in the representation of labels, which allows to work on many uncertainty frameworks, such as fuzzy sets, possibilities or even classical probabilities. The datasets[7] are free and available, they also contain the raw user answers for any type of need.[8]

References

[1] Howe and Jeff, "The rise of crowdsourcing," *Wired*, vol. 14, 2006.

[7]Link to the datasets: https://data.mendeley.com/datasets/4hz3wx6wm5

[8]Pictures of dog datasets are for the majority resized from ImageNet [17] with some personal imports. It is inspired by Stanford Dogs dataset [18]. For bird datasets, some pictures are ours, but we also thank the contributors of Pixabay and Wikimedia. All aliases are listed in the repository. This work is funded by the Brittany region and the Côtes-d'Armor department.

[2] E. Schenk and C. Guittard, "Crowdsourcing: What can be outsourced to the crowd, and why ?" 2009.

[3] A. Felstiner, "Working the crowd: Employment and labor law in the crowdsourcing industry," *Berkeley Journal of Employment & Labor Law*, vol. 32, pp. 142–204, 2011.

[4] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[5] C. Thierry, A. Hoarau, A. Martin, J.-C. Dubois, and Y. Le Gall, "Real bird dataset with imprecise and uncertain values," in *7th International Conference on Belief Functions*, 2022.

[6] P. Smets, *Imperfect Information: Imprecision and Uncertainty*. Boston, MA: Springer US, 1997, pp. 225–254.

[7] C. Thierry, A. Martin, J.-C. Dubois, and Y. Le Gall, "Validation of Smets' hypothesis in the crowdsourcing environment," in *6th International Conference on Belief Functions*, Shanghai, China, Oct. 2021.

[8] L. Schmarje, J. Brünger, M. Santarossa, S.-M. Schröder, R. Kiko, and R. Koch, "Fuzzy overclustering: Semi-supervised classification of fuzzy labels with overclustering and inverse cross-entropy," *Sensors*, 2021.

[9] L. Schmarje, C. Zelenka, U. Geisen, C.-C. Glüer, and R. Koch, "2d and 3d segmentation of uncertain local collagen fiber orientations in SHG microscopy," in *Lecture Notes in Computer Science*. Springer International Publishing, 2019, pp. 374–386.

[10] A. P. Dempster, "Upper and Lower Probabilities Induced by a Multivalued Mapping," *The Annals of Mathematical Statistics*, 1967.

[11] G. Shafer, *A Mathematical Theory of Evidence*. Princeton: Princeton University Press, 1976.

[12] T. Denœux, "The cautious rule of combination for belief functions and some extensions," *2006 9th International Conference on Information Fusion, FUSION*, pp. 1 – 8, 08 2006.

[13] Z. Elouedi, K. Mellouli, and P. Smets, "Belief decision trees: theoretical foundations," *International Journal of Approximate Reasoning*, 2001.

[14] T. Denœux, "A k-nearest neighbor classification rule based on dempster-shafer theory," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 219, 1995.

[15] T. Denoeux and M. Bjanger, "Induction of decision trees from partially classified data using belief functions," *Systems, Man, and Cybernetics, 2000 IEEE International Conference*, vol. 4, pp. 2923–2928, 2000.

[16] A. Hoarau, A. Martin, J.-C. Dubois, and Y. Le Gall, "Imperfect labels with belief functions for active learning," in *Belief Functions: Theory and Applications*. Springer International Publishing, 2022.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[18] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.