Real bird dataset with imprecise and uncertain values

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

Real bird dataset with imprecise and uncertain values

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

Univ Rennes, CNRS, IRISA, DRUID, France

Abstract. The theory of belief functions allows the fusion of imperfect data from different sources. Unfortunately, few real, imprecise and uncertain datasets exist to test approaches using belief functions. We have built real birds datasets thanks to the collection of numerous human contributions that we make available to the scientific community. The interest of our datasets is that they are made of human contributions, thus the information is therefore naturally uncertain and imprecise. These imperfections are given directly by the persons. This article presents the data and their collection through crowdsourcing and how to obtain belief functions from the data.

Keywords: Datasets · Imprecise · Uncertain

1 Introduction

The theory of belief functions allows for uncertainty and imprecision in the data. However, there are very few real datasets available to consider belief functions. [4] and [1] work on imprecise and uncertain real data that the authors have collected but these data are not made available to the community. Similarly, [3] proposes an MCQ that allows students to give imprecise and uncertain answers that can be modeled with belief functions. However, the experimental data are not reported. It was important to build real datasets to evaluate proposed methods in real context [9]. To do so, we collected human contributions from crowdsourcing campaigns, as human contributions are uncertain and imprecise information.

Crowdsourcing is the outsourcing of tasks to a crowd of contributors on a platform dedicated to the domain [5]. The tasks that can be achieved through crowdsourcing are very diverse. In this paper, we presented to the contributors a picture of a bird and asked them to identify the bird from a list of proposed names. We use interfaces that allow us to collect imprecise and/or uncertain responses. We conducted six crowdsourcing campaigns for bird photo annotation. For all these campaigns the contributor had to give his certainty in his answer. Two of them are only composed of precise contributions. For the four other campaigns, the contributor can be imprecise and choose more bird names. For two of the imprecise campaigns, after the contributor has given his answer he is offered to enlarge or restrict his selection consequently.

The rest of the paper is as follows, section 2 introduces the belief functions. Section 3 reviews the crowdsourcing campaigns and section 4 presents the datasets. We propose examples of modelisation thanks to the belief function section 5. Section 6 concludes the paper.

2 Belief functions

The theory of belief functions, also called Dempster-Shafer theory [2,8], is used in this study in order to model both data imprecision and uncertainty.

One considers $\Omega = \{r_1, \ldots, r_M\}$ the frame of discernment for M exclusive and exhaustive hypotheses. In this paper, Ω represents all possible bird species of a given photo among M bird species. 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.

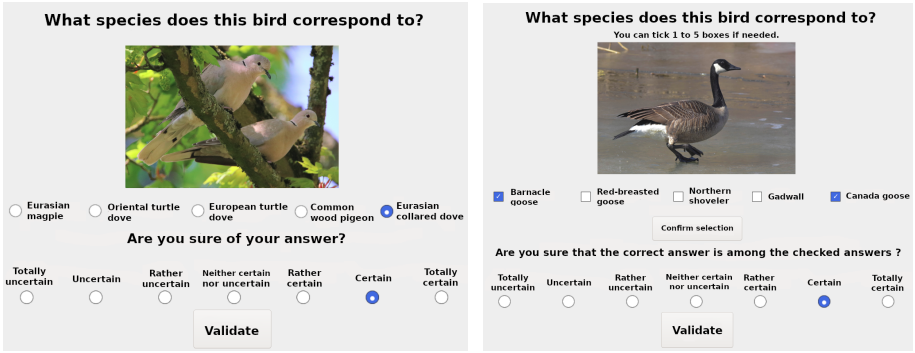
$$\begin{aligned} & m : 2^\Omega \rightarrow [0,1] \\ & \sum_{A \in 2^\Omega} m(A) = 1 \end{aligned} \tag{1}$$

Focal element: An element of 2^Ω with a non-null mass.
Simple support mass function: Only has two focal elements, and one of them is the frame of discernment Ω.
Consonant mass function: Each focal element is nested.

3 Crowdsourcing Campaign

The main objective for these campaigns is always the same, a photo of a bird is presented to the contributor with a set of species names (including the good answer) and he has to select the right answer. The Wirk platform (Crowdpanel[1]) is used to realize the crowdsourcing campaigns. As the users of the platform live in France, the birds used for the campaigns are all of species visible in metropolitan France. For all the campaigns, the contributor has to give his answer, then specify his certainty according to the following Likert scale: "Totally uncertain","Uncertain", "Rather uncertain", "Neutral", "Rather certain", "Certain", "Totally certain". We explained to the contributors that there is no penalty for being uncertain and/or imprecise in their answers. After having given his answer and his certainty, he can validate his contribution in order to move on to the next question.

[1] https://crowdpanel.io/ (15/04/2022)



(a) multi_birds_precise (b) multi_birds_imprecise

Fig. 1. Interfaces used for crowdsourcing campaigns multi_birds

3.1 Campaigns multi_birds

For these two campaigns, five bird names are proposed to the contributor. The names change from one question to another and a bird species is presented only once. We have tried to introduce different levels of difficulty in the questions. For example, for a difficult question, a photo of an eagle is presented to the contributor and the five answer items are different species of eagles. Conversely, for a simpler question, a photo of a gull is presented to the contributor and the four other answers are names of duck species. For a single photo, responses were presented in random order to each contributor to avoid selection bias. In addition, the questions were also asked in a random order, so that when a contributor c_1 answers a question q_i, c_2 answers q_j. These crowdsourcing campaigns include 3 attention questions for which the contributor is asked to give the same answer as the one given in the previous question.

multi_birds_precise The interface used for this task is given figure 1.(a). Participants have to provide a precise answer by selecting a single bird name, and a self-assessment of their certainty in this answer.

multi_birds_imprecise For this task the contributor can be imprecise and select up to all of the bird names offered. The interface is given figure 1.(b). The contributors first must give his answers, validate it and then he is asked to give his certainty in this answer.

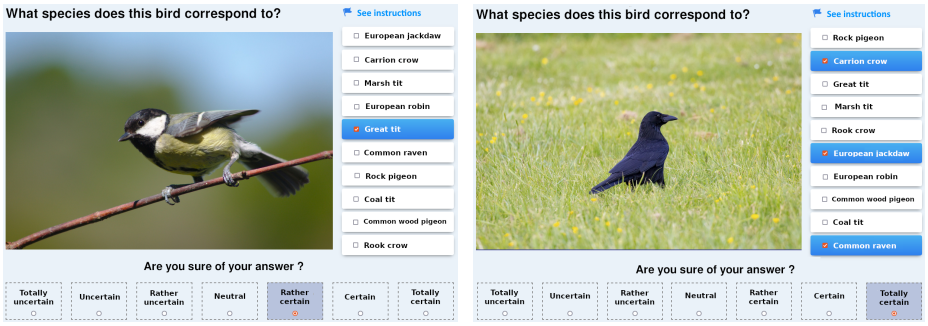
For both campaigns the crowds are composed of 100 contributors, each one must annotate 50 photos, for a total of 5000 contributions for each campaign. A contributor allowed to do the first campaign cannot participate in the second.

3.2 Campaigns 10_birds

For these campaigns, ten bird species are selected and proposed as response elements to the contributors. In order to observe the contributor's ability to be imprecise in case of hesitation, the ten birds presented are composed of subgroups

Table 1. The ten bird species used in the 10_birds campaigns group by family

Muscicapidae	Columbidae	Paridae	Corvidae
European robin	Common wood pigeon	Great tit	European jackdaw
	Rock pigeon	Marsh tit	Carrion crow
		Coal tit	Common raven
			Rook crow



(a) precise (b) imprecise

Fig. 2. Interfaces used for crowdsourcing campaigns 10_birds precise and imprecise

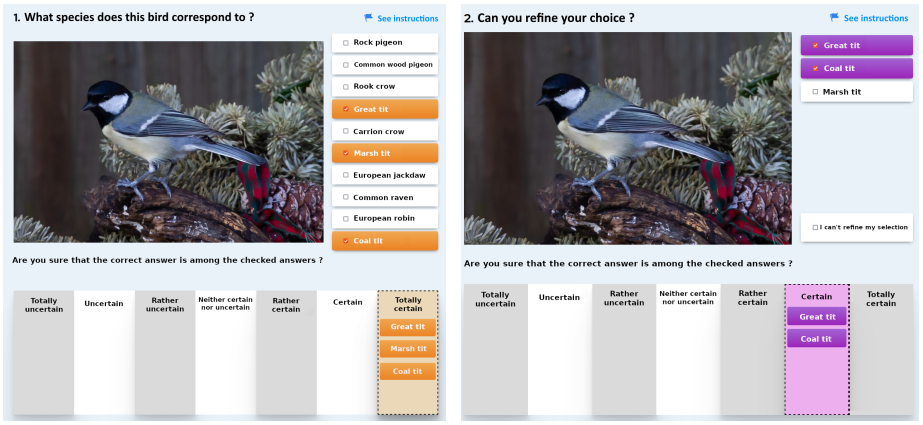
from the same bird family given table 1. The bird names are presented to each contributor in a different order to avoid selection bias. This ordering of names is nevertheless fixed for a contributor throughout the campaign. Such as the campaigns multi_birds, the questions are asked in a random order. The same scale is used for certainty and 3 attention questions are also asked. The contributor is no longer required to validate his answer before he can give his certainty.

10_birds_precise The contributor should select from the interface figure 2(a) a unique bird name and then give his certainty about it.

10_birds_imprecise The contributor can choose thanks to the interface figure 2(b) 1 to a maximum of 5 answers from the ten provided bird names. We impose a maximum number of answers to 5 because we admit that if the contributor hesitates it is between names of birds of the same family. He should not hesitate between a pigeon and a chickadee for example. We have chosen to offer the crowd a maximum selection of 5 names because we do not want to introduce a bias and encourage him to choose exactly the 4 corvidaes in case of hesitation.

10_birds_iterative This campaign is called iterative because the contributor is asked to expand or refine the contribution they have entered. To do so, in a first step the contributor answer the question as shown figure 3(a) and then:

- If he is precise but not "totally certain" of his answer, he is offered to expand his selection if he feels the need. In this case, the first selected answer is kept in step 2 and he can complete it by selecting new names.
- If he is imprecise in his contribution, he is asked in a second step if he is able to restrict his choice of answer while giving his new certainty as in the



(a) iterative step 1 (b) iterative step 2

Fig. 3. Interfaces used for the campaigns 10_birds iterative and machine learning

example figure 3(b). When he is offered to restrict his selection, only the previously chosen answer elements are proposed again.

These interactions with the contributor increased the number of responses collected, and therefore the time spent soliciting the contributor.

10_birds_machine_learning For classification problems, a larger number of observations is often required. This campaign provides imperfectly labeled observations that can be used for classification problems and more generally in machine learning. This is a similar campaign to 10_birds_iterative with more photos per class and fewer responses per photo. A number of 20 photos per bird species is used instead of 5. A total of 200 photos separated into 10 species are then labelled. Contributors are only invited to give 20 answers, 2 randomly selected per species of bird. No attention questions are asked in this campaign.

Each crowdsourcing campaign required a crowd of 50 contributors. As with the other campaigns, a contributor who has participated in one experiment cannot participate in another. For each of the ten bird that make up the proposed answer set, a contributor is presented with 5 photos of a bird, so that he answers 50 questions. Thus 2500 data are collected for the experiments 10_birds_precise, with precise answers, and 10_birds_imprecise, for which the contributor can choose up to five answers. There is 2990 data collected for 10_birds_iterative because for this experiment, as there are 2550 first step answers and 440 second step answers. Finally 1515 data are collected for 10_birds_machine_learning, with 1040 first step answers and 475 second step answers.

All crowdsourcing campaigns are summarized in Table 2 which includes the name of the campaign, whether the data collected is accurate or inaccurate, the number of contributors and the total number of contributions collected.

Table 2. Summary of the crowdsourcing campaigns conducted, the certainty is asked for all the answers

Campaign	Answers	Crowd size	Number of answer
multi_birds_precise	Precise	100	5000
multi_birds_imprecise	Imprecise	100	5000
10_birds_precise	Precise	50	2500
10_birds_imprecise	Imprecise	50	2500
10_birds_iterative	Imprecise	51	2990
10_birds_machine_learning	Imprecise	52	1515

4 Details of the data sets

The datasets are made available to the community on the INRIA git account https://gitlab.inria.fr/cthierry/imprecise_uncertain_dataset. The repository is structured as follows: a folder is associated with each crowdsourcing campaign presented above, there is also a csv file named "answers_multi_birds.csv". This file is used as the answer set propose to the contributor of the multi_birds campaign. It includes the following variable:

- photo: the number of the bird photo display by the interface
- goodanswer: the true bird name of the photo
- answ1, answ2, answ3, answ4: other bird names propose as answer elements to the contributor in addition of the true bird name
- difficulty: an hypothesis of the difficulty of the question according to the author. Values range from 1 (easy question) to 5 (difficult question). In fact the difficulty observed is not correlated to those one supposed by the author.

Thus, it is possible to use the file "answers_multi_birds.csv" to construct a frame of discernment $\Omega_q = \{goodanswer, answ1, answ2, answ3, answ4\}$ for each question q of the multi_birds campaigns. In the files named after the crowdsourcing campaigns there are several files providing different information.

Data This csv file includes the contribution from the crowd for the bird annotation. The contribution includes therefore a selection of bird names and a certainty associated. The file includes:

- log_id: indicates the line of the file
- user: unique user ID for a contributor
- currenttrial: number of the question asked to the contributor
- img: number of the photos shown to the contributor
- goodanswer: the true name of the bird to identify
- answer: the set of bird names selected by the contributor (a unique answer for the precise crowdsourcing campaign)
- answerhistory: set of values checked/unchecked by the contributor to answer the question
- isgoodanswer: boolean indicated if the true bird name is include in the answer set given by the contributor

- certitude: certainty given by the contributor to express his confidence in his bird names selection
- certitudehistory: history of the certainty values for the answer
- timestamp: time recorded by the interface
- time: time of the contributor's answer to the question

Some variables such as user are common to several files. For all the files, the contributor certainty ranges from 1 (totally uncertain) to 7 (totally certain).

For the campaign multi_birds_imprecise 57.04% of the data includes in the csv file are imprecise. And for the 10_birds campaigns we have the following results: imprecise 55.64%, iterative 45.32% and machine_learning 58.22%. Contributors have made good use of the opportunity to be imprecise when possible. On average, when contributors are imprecise they choose two answers.

Attention In crowdsourcing campaigns, attention questions are asked to the contributor in order to ensure its seriousness. This csv file includes the answers to the three attention questions. These questions consist in asking the contributor a previous question in order to get him to give exactly the same contribution. The variables included in the file are the following:

- log_id, user, currenttrial, answer, certitude, timestamp, time
- attention_answer: contributor's answer to the attention question
- answerhistory: set of values checked/unchecked by the contributor to answer the attention question
- certainty: certainty of the first selected answer set
- certaintyhistory: history of checked/unchecked certainty values
- issamecertitude: boolean indicating if the certainty given to the attention question is identical to the certainty given to the initial question
- issameanswer: boolean indicating if the set of bird names selected at the attention question is identical to the set initially selected

Event This file records the principal events of the platform named event_type: the connection to the platform (start), the beginning of the crowdsourcing campaign (start_xp), the ending question (questions) and the end of the campaign (finish). It is possible that some contributors have started a crowdsourcing campaign without having finished it, we have only a part of the answers for them. To sort out the contributors (users) to be selected we recommend using the event.csv or question.csv files described below to select the data of users who have reached the final question phase and/or the finish event. When we talk about the number of responses, it is only for contributors who have completed the entire campaign. This file also includes the variables log_id, user and time which gives the date and time when the event took place.

Queries At the end of the different crowdsourcing campaigns, a questionnaire is sent to the contributors to get their feedback. This questionnaire varies between campaigns. These files include the answers at the end of the campaign.

Iteration This csv file is present in the 10_birds_machine_learning and 10_birds_iterative folders because it includes the contributors' answers when they expand or specify their answer in the second stage of questioning of these campaigns. The next values are included into the file:

Table 3. Number of precise and then imprecise contributions ($|X_1| = 1$ and $|X_2| > 1$) and imprecise then less imprecise ($|X_1| > |X_2|$).

Campaigns	Size of the dataset	Data subset	Size of the of the data subset
10_birds_iterative	440	$\|X_1\| = 1$ and $\|X_2\| > 1$	88
		$\|X_1\| > 1$ and $\|X_1\| > \|X_2\|$	352
10_birds_machine_learning	475	$\|X_1\| = 1$ and $\|X_2\| > 1$	57
		$\|X_1\| > 1$ and $\|X_1\| > \|X_2\|$	418

- log_id, user
- trial: equals to currenttrial value of the same contributor (user) in data.csv
- new_answer: new answer given by the contributor
- new_certitude: new certainty given by the contributor
- cant_answer: boolean that takes the value 1 if the contributor cannot modify (refine or enlarge) his answer
- isImprecis: boolean which takes the value 1 if in his first answer the contributor is imprecise (*i.e.* he chooses several bird names)
- aHistory: equivalent to answerhistory
- cHistory: equivalent to certaintyhistory

The file for the campaign 10_birds_iterative includes 1527 rows but for the majority of them, the contributor did not modify his answer (cant_answer=1). Indeed, for this campaign only 440 responses were modified which represents 17% of the first step answers. More contributors edited their answer for the 10_birds_machine_learning campaign, 475 responses were modified i.e. 46% of the dataset. Thanks to the joint use of this file and data.csv it is possible to build 440 consonant mass functions.

For campaigns with iteration we call X_1 the first set of names given for a photo. The values of X_1 (answer) are present in the data.csv file with the associated certainty. When the contributor is proposed to modify his contribution, the new name selection X_2 (new_answer) and the new certainty (new_certitude) are registered in the iteration.csv file.

For the campaign 10_birds_iterative there are 461 entries in the data.csv file for which the contributor first selected a single answer $|X_1| = 1$, and then was offered to expand his selection so that $|X_2| > 1$. Of these 461 times when the contributor is offered to be imprecise, there are only 88 times when a second answer X_2 is given. Similarly, there are a total of 1066 times when the contributor fills in an imprecise answer, $|X_1| > 1$, and is offered to narrow his selection so that $|X_1| > |X_2|$, a total of 352 contributions report a change in answer.

During the 10_birds_machine_learning campaign contributors also tend to give a second answer more precise than the first one, with 418 responses on 475 iterations, against 57 precise answers at firt step and less precise at second step. Furthermore, 389 responses were listed as totally certain and 186 were listed as totally certain and precise. Among those contributors who were certain and precise, 91% hold the real answer. Of all the answers, 33 were listed as

inconsistent, which means that a contributor gave an answer including different bird families, these 33 answers were generated by 15 different contributors.

The following section presents examples of modeling with mass functions.

5 Belief functions from the data

We propose a data modeling by simple support and consonant mass functions.

Simple support mass function This function can be computed for the answer values and certainty of the six campaigns. Given a question q, the set of answers associated to q compose the framework of discernment $\Omega_q = \{r_1, \ldots, r_K\}$. The question being closed, we consider the closed world. The contributor c answers the question q by the contribution $X \in 2^{\Omega_q}$, which can be imprecise, and to which he associates a certainty of value $certainty \in [1, 7]$ which is transformed into a mass $\omega_{cq} \in [0, 1]$ according to the equation:

$$\omega_{cq} = \frac{certainty - 1}{certainty_{max} - 1} \tag{2}$$

For the crowdsourcing campaign introduce in this paper $certainty_{max} = 7$. A mass function with simple support ($X^{\omega_{cq}}$) can be obtained from the contribution:

$$\begin{cases} m_{cq}^{\Omega_q}(X) = \omega_{cq} \text{ with } X \in 2^{\Omega_q} \setminus \Omega_q \\ m_{cq}^{\Omega_q}(\Omega_q) = 1 - \omega_{cq} \end{cases} \tag{3}$$

Consonant mass function During the 10_birds_machine_learning and 10_birds_iterative campaigns, the same question q can be asked twice to the contributor c who can then enlarge or specify his first answer X_1 by a second answer X_2 if he wishes. Let Ω_q be the set of proposed answers and $X_1, X_2 \in 2_q^{\Omega}$. If the first answer of the contributor X_1 is precise and he widens his second answer X_2 then $X_1 \subset X_2$, and conversely if X_1 is more imprecise than X_2 then $X_2 \subset X_1$. At the time of his first selection X_1, the contributor informs a degree of certainty of numerical value $\omega_{cq1} \in [0, 1]$ compute thanks to equation (2). If he chooses to fill in a second answer X_2 he must indicate his new certainty whose numerical value is noted $\omega_{cq2} \in [0, 1]$. If the contributor is not asked to modify his selection or if he does not wish to do so, the contribution is modeled by a simple support mass function. In the case where the contributor changes its response X_1 to the response X_2, with $X_1 \subset X_2$, then the contribution can be modeled by a consonant mass function:

$$\begin{cases} m_{cq}^{\Omega_q}(X_1) = \delta_1 * \omega_{cq1} \\ m_{cq}^{\Omega_q}(X_2) = \delta_2 * \omega_{cq2} \\ m_{cq}^{\Omega_q}(\Omega) = 1 - \delta_1 * \omega_{cq1} - \delta_2 * \omega_{cq2} \end{cases} \tag{4}$$

In equation (4), the coefficients δ_1 and δ_2 ensure that the mass function belongs to the interval $[0, 1]$, thus: $\delta_1 + \delta_2 = 1$. If we want to give more importance to the first contribution X_1 rather than to the second contribution X_2 then we must

choose δ_1 such that $\delta_1 > \delta_2$. Another way to combine the two mass functions from the two iterative responses is to use a combination rule that does not require the assumption of source independence.

We have proposed a modeling of some data by belief functions but it is possible to go further by using them for example to estimate the expertise of the contributor as do [7]. The data can also be used to compare a probabilistic approach to belief functions [6].

6 Conclusion

This paper presents some real credal datasets created through crowdsourcing campaigns for bird photo annotation. To constitute these datasets six crowdsourcing campaigns have been realized. In these six campaigns, the contributor is asked to give his certainty in his answer. For two campaigns the contributor is forced to choose a single bird name as an answer, these data are therefore precise and potentially uncertain. For the other four campaigns the contributor had the possibility to be imprecise in case of hesitation on his answer, these data are imprecise and/or uncertain. For these six crowdsourcing campaigns it is possible to model the contributions by simple support mass functions. Finally, for two of the four imprecise campaigns, the contributor is asked to modify the answer already given by clarifying or expanding it. Thanks to these two campaigns it is possible to model the contributions by consonant mass functions.

References

1. Abassi, L., Boukhris, I.: A worker clustering-based approach of label aggregation under the belief function theory. Applied Intelligence pp. 1–10 (2018)
2. Dempster, A.P.: Upper and Lower Probabilities Induced by a Multivalued Mapping. The Annals of Mathematical Statistics **38**(2), 325 – 339 (1967)
3. Diaz, J., Rifqi, M., Bouchon-Meunier, B., Jhean-Larose, S., Denhiére, G.: Imperfect answers in multiple choice questionnaires. In: European Conference on Technology Enhanced Learning. pp. 144–154. Springer (2008)
4. Dubois, J.C., Gros, L., Kharoune, M., Le Gall, Y., Martin, A., Miklós, Z., Ouni, H.: Measuring the Expertise of Workers for Crowdsourcing Applications. In: Advances in Knowledge Discovery and Management, pp. 139–157 (Jun 2019)
5. Howe, J.: The rise of crowdsourcing. Wired Magazine (2006)
6. Koulougli, D., Hadjali, A., Rassoul, I.: Handling query answering in crowdsourcing systems: A belief function-based approach. In: Fuzzy Information Processing Society (NAFIPS), 2016 Annual Conference of the North American. pp. 1–6. IEEE (2016)
7. Martin, A.B.R.M.K.Z.M.A.: Characterization of experts in crowdsourcing platforms. Belief Functions: Theory and Applications. **9861** (2016)
8. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
9. Thierry, C., Martin, A., Dubois, J.C., Le Gall, Y.: Validation of Smets' hypothesis in the crowdsourcing environment. In: 6th International Conference on Belief Functions. Shanghai, China (Oct 2021)