Imperfect Labels with Belief Functions for Active Learning

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

Imperfect Labels with Belief Functions for Active Learning

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

Univ Rennes, CNRS, IRISA, DRUID, France

Abstract. Classification is used to predict classes by extracting information from labeled data. But sometimes the collected data is imperfect, as in crowdsourcing where users have partial knowledge and may answer with uncertainty or imprecision. This paper offers a way to deal with uncertain and imprecise labeled data using Dempster-Shafer theory and active learning. An evidential version of K-NN that classifies a new example by observing its neighbors was earlier introduced. We propose to couple this approach with active learning, where the model uses only a fraction of the labeled data, and to compare it with non-evidential models. A new computable parameter for EK-NN is introduced, allowing the model to be both compatible with imperfectly labeled data and equivalent to its first version in the case of perfectly labeled data. This method increases the complexity but provides a way to work with imperfectly labeled data with efficient results and reduced labeling costs when coupled with active learning. We have conducted tests on real data imperfectly labeled during crowdsourcing campaigns.

Keywords: Belief Functions · Imperfect Labels · Active Learning.

1 Introduction

In supervised classification, where the aim is to find the class of an observation, one still works largely with hard labels *i.e.* if a label exists for an observation, this label is defined in a categorical way. The labeling process often is carried out by humans [7, 10]; without making any difference between a label given by someone who has hesitated for a long time and someone who has no doubt.

Using hard labels might be convenient for many machine learning and deep learning problems but is never completely representative of the reality. Imperfection, on the other hand, can help us fill in this lack of information. It can be represented by many criteria but only uncertainty and imprecision will be discussed in this paper. Ignorance is then derived from imprecision. Such information can be modeled with the theory of belief functions, introduced in [1, 12]. This paper proposes to compare a non-evidential model with its evidential version and to observe the impact of imperfect labeling on classification. The widely used non-parametric model K-NN [5] will then be compared with EK-NN, an evidential version presented in [2–4]. A new parameter will be proposed for EK-NN to work with imperfectly labeled data and to maintain equivalence with the

original model. In a context where data labeling is not only imperfect but also expensive, active learning [11] is particularly interesting. Indeed, a small volume of labeled data is sufficient to obtain good performance. Very little research has been done to couple belief functions with active learning. The main difference with [14] is the use of an imperfectly labeled data set instead of using only noise. The plan of the paper is as follows. Section 2 reviews the theory of belief functions, K-NN and EK-NN algorithms and then ends with an overview of active learning. Section 3 describes the proposed method, and the contribution concerning the parameters of EK-NN. A new credibilist dataset is also presented. Experiments on datasets composed of noisy real data and imperfectly labeled data are discussed in section 4. Finally, Section 5 concludes the article.

2 Background

2.1 Reminder on Belief Functions

The theory of belief functions, also called Dempster-Shafer theory [1,12], is used in this study in order to model both imprecision and uncertainty.

One considers $\Omega = \{\omega_1, \ldots, \omega_M\}$ 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 (BBA) 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.

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

Each subset $A \in 2^\Omega$ such as $m(A) > 0$ is called a *focal element* of m. If $m(A) = 1 - \delta$ and $m(\Omega) = \delta$ with $A \in 2^\Omega \backslash \emptyset$ and $\delta \in [0,1]$, m is called a *simple support mass function*.

A source might not be trustworthy, a discounting coefficient α is then introduced to transfer some belief into Ω, also called the ignorance, such that:

$$\begin{cases} m_\alpha(A) = \alpha m(A)\ , \quad \forall A \in 2^\Omega, A \neq \Omega, \\ m_\alpha(\Omega) = 1 - \alpha(1 - m(\Omega)), \end{cases} \tag{2}$$

where m_α is the new discounted mass.

The normalized conjunctive combination of Basic Belief Assignments (BBAs) m^j derived from $\mathcal{N}$ sources is given by:

$$\begin{cases} m(A) = \dfrac{1}{1-\kappa} \displaystyle\sum_{B_1 \cap \ldots \cap B_\mathcal{N} = A} \prod_{j=1}^{\mathcal{N}} m^j(B_j) \ \text{ if } A \neq \emptyset, \\ m(\emptyset) = 0, \end{cases} \tag{3}$$

with $\emptyset$ the empty set and:

$$\kappa = \sum_{B_1 \cap \ldots \cap B_{\mathcal{N}} = \emptyset} \prod_{j=1}^{\mathcal{N}} m^j(B_j). \tag{4}$$

On decision level, the pignistic probability $BetP$ helps decision making on singletons:

$$BetP(\omega) = \sum_{A \in 2^{\Omega},\, \omega \in A} \frac{m(A)}{|A|}. \tag{5}$$

2.2 *K*-Nearest Neighbors

When dealing with perfectly labeled data, a non-parametric discrimination model known as the K-Nearest Neighbors (K-NN) is introduced in [6]. This is a popular classification model in which the label of an incoming sample is predicted according to its K nearest neighbors. The main drawback of this algorithm is that it assumes that there are close neighbors of the incoming sample. It is then proposed in [5] a distance-weighted K-NN where each neighbor is weighted according to its closeness to the incoming sample.

2.3 E*K*-NN

An evidential version of K-NN is introduced in [3], this Evidential K-Nearest Neighbors (EK-NN) uses belief functions to assign a label to a new sample. It is presented in the original paper as working with perfectly labeled data, but some work has subsequently been done to make this algorithm work with imperfectly labeled data. A version of EK-NN [4] is proposed with data labeled with possibility theory and [2] allows to calculate the parameters when dealing with imperfectly labeled data coupled with the theory of belief functions. However, it then loses the equivalence with the previous model in the particular case of perfectly labeled data.

2.4 Active Learning

Imperfect labels can be modeled by belief functions, and EK-NN can be a tool for learning from imperfectly labeled data, but we are also interested in reducing the number of labeled instances. Active learning [11] is a part of machine learning where the learner can choose which observation to label in order to work with only a fraction of the labeled data to reduce the labeling cost. Observations are called *Instances*, the act of requesting for the label of an instance is a *Query* and the entity giving its label to an instance is called the *Oracle*.

The difficulty is therefore to determine which instances should be labeled first. This process is called *Sampling*, the best known being Random Sampling where queries to the Oracle are made on random instances. Uncertainty Sampling, on the other hand, aims to perform a query on the sample for which the model is the least certain.

3 Classification of imperfectly labeled data with E*K*-NN and active learning

Let $\mathcal{X}$ be a P features collection of N samples such as $\mathcal{X} = \{x^n = (x_1^n, \ldots, x_P^n) \mid n = 1, \ldots, N\}$, and Ω a set of M classes as $\Omega = \{\omega_1, \ldots, \omega_M\}$. Let $d^{s,i}$ be the distance between x^s and x^i with x^s an incoming sample to be classified using the information contained in the training set and x^i one of its K nearest neighbors. Classifying x^s means assigning it one class in Ω. Let Φ^s be the set of the K-nearest neighbors of x^s in $\mathcal{X}$ and m^i the BBA associated to x^i.

3.1 E*K*-NN for imperfectly labeled data

In [3], the author introduces an equation of the BBA between an unclassified sample x^s and a neighbor x^i when it comes to imperfectly labeled data. This section results from the following proposition. If x^s is a sample to be classified, one's belief about the class of x^s induced by knowing that $x^i \in \Phi^s$ can be represented by a basic belief assignment $m^{s,i}$ deduced from m^i and $d^{s,i}$:

$$\begin{aligned} m^{s,i}(A) &= \alpha_0 \phi(d^{s,i}) m^i(A), \\ m^{s,i}(\Omega) &= 1 - \sum_{A \in 2^\Omega \setminus \Omega} m^{s,i}(A), \end{aligned} \tag{6}$$

with ϕ a monotonically decreasing function and:

$$\begin{aligned} &0 < \alpha_0 < 1, \\ &\phi(0) = 1, \\ &\lim_{d \to \infty} \phi(d) = 0. \end{aligned} \tag{7}$$

As a decreasing function ϕ, [3] suggests to choose:

$$\phi(d) = e^{-\gamma d^\beta}, \tag{8}$$

with $\gamma > 0$ and $\beta \in \{1, 2, \ldots\}$ possibly fixed to a small value. When ϕ is first introduced, it depends on γ_q with ω_q the class of x^i and there are as many γ_q as different classes. As each point x^i no longer has a unique label since we are using imperfectly labeled data, γ_q cannot be calculated. This specificity forces the model to differ from a model using hard labeled data. It is discussed in section 3.2.

Each BBA is now combined using (3):

$$\bar{m}^s(A) = \sum_{B_1 \cap \ldots \cap B_K = A} \prod_{x^i \in \Phi^s} \alpha_0 \phi(d^{s,i}) m^i(B_i), \forall A \in 2^\Omega. \tag{9}$$

Considering the closed world, the mass of the empty set must be forced to be null. The new normalized combined BBA, denoted m^s is obtained as:

$$\begin{cases} m^s(A) = \dfrac{1}{1-\kappa} \bar{m}^s(A), & A \neq \emptyset, \\ m^s(\emptyset) = 0. \end{cases} \tag{10}$$

with κ the fusion inconsistency given at equation (4). Each new sample is then classified by maximizing the pignistic probability.

3.2 Parameters Optimization and γ_i-EKNN

This part deals with the calculation of the parameters of the model. They are: K, α_0, γ and β. The number K of nearest neighbors can be optimized identically to K-NN, using for example cross-validation. Furthermore, the use of variable size datasets within active learning has an impact on the optimal K. From preliminary experimental results (not given here), the parameter α_0 is set to 0.8, but might depend on the knowledge of the sources, which modifies the results very slightly; $\beta = 2$ gave satisfying results with little impact when changed. When dealing with imperfectly labeled data, the use of one γ parameter per class becomes meaningless, as there are no longer any classes but only samples with BBAs that more or less belong to a class. Several options have been proposed in [2–4] and compared in [2].

- In its first version [3], here renamed γ_q-EKNN, the model is presented with a γ_q parameter depending on the class ω_q of the neighbor x^i. The computable formula given for γ_q is $1/\mathsf{d}_q^\beta$ with d_q the mean distance between two training vectors of the same class.
- A one γ version of the model [4], γ-EKNN, is later presented in a possibilistic environment and suitable for imperfectly labeled data. The use of a single γ parameter leads to the loss of equivalence with the initial model.
- Finally, a contextual-discounting based model [2] with M learnable γ_q is introduced, and will be referred to as CD-EKNN.

In this paper, we propose γ_i-EKNN, a version with K computable γ_i parameters, allowing to recover the equivalence, both theoretical and practical, with the original model in the case of perfectly labeled data. To maintain the equivalence with the model introduced in [3] when dealing with perfectly labeled data, the proposition of using one γ for each neighbor according to their similarity is made. When it comes to imperfectly labeled data, γ is calculated in relation to the distance with the other samples and according to its resemblance with Jousselme distance introduced in [8]. The closer we get to perfect labeling, the closer we get to one γ per class:

$$\gamma_i = \frac{1}{\mathsf{d}_i^\beta}, \qquad \mathsf{d}_i = \frac{\sum_{\nu=0}^{N}\sum_{\mu=0}^{N}(1-d_J^{i,\nu})(1-d_J^{i,\mu})d^{\nu,\mu}}{[\sum_{\nu=0}^{N}(1-d_J^{i,\nu})]^2 - \sum_{\nu=0}^{N}(1-d_J^{i,\nu})^2}, \tag{11}$$

with N the total number of samples and $d_J^{i,\nu}$ Jousselme's distance between m^i and m^ν.

In order to study the relevance of using imperfect labels by comparing an eviendential and a non-evidential model, γ_i-EKNN will be used, at the cost of its complexity, as it maintains equivalence with the orginal model.

3.3 Labeling with Uncertainty and Imprecision

In order to work with imperfectly labeled data, we obtained a dataset from crowdsourcing campaigns using the model and the materials developed in [13].

Credal Bird-10 is a dataset composed of 200 pictures of birds imperfectly labeled. Each of these images belongs to a class corresponding to one of the 10 species of birds evenly distributed on the dataset. During crowdsourcing campaigns, the pictures are displayed and participants can choose multiple corresponding classes as well as the belief they have in their responses. The resulting dataset is a combination of pictures associated with BBAs, refer to [13] for construction of the dataset. When using non-evidential models, the class maximizing the pignistic probability is then chosen as the perfect label. Two datasets have been obtained, one on the Irisa laboratory (*Credal Bird-10 irisa*) and one on a non-specific crowd of paid contributors (*Credal Bird-10 public*).

Example for a picture of red/green/blue pixels corresponding to a marsh tit with y_1 the vector on 2^Ω which is the BBA describing its imperfect label:

	y_1
$m(\{Marsh\ tit\})$	0.2
$m(\{Marsh\ tit,\ Great\ tit\})$	0.5
$m(\Omega)$	0.3

4 Experiments

The following section presents several procedures for implementing the method. The interest is to show that allowing a source to provide imperfectly labeled data may be more realistic than perfectly labeled data and therefore yield better results. For each experiment 20% of the dataset is used as a test set and the remaining as a training set. The experiment of section 4.1 is a brief comparison between the approaches discussed in section 3.2. The experiments given in sections 4.2 and 4.3 are coupled to active learning in order to avoid expensive labeling. A comparison between K-NN and its evidential version is made to study the relevance of using imperfectly labeled data, other models are added for a general overview.

4.1 Different approaches for γ parameter

A comparison is made in table 1 between K-NN and the approaches presented in section 3.2. They are used with a K nearest neighbors value equal to 7 and the result is a mean accuracy over 100 iterations. The distance weighted K-NN is compared to the original version γ_q-EKNN, to the unique gamma γ-EKNN version using $\gamma = 1/\mathsf{d}^\beta$, with d the mean distance between two training vectors, and to the proposed γ_i-EKNN. Datasets are split into two categories: perfectly

labeled (Iris, Wine and Breast Cancer)[1] and imperfectly labeled (Credal Bird-10 public). The 95% confidence interval[2] is also given.

As it can be observed in table 1, there is an equivalence between γ_q-EKNN and the proposed γ_i-EKNN on perfectly labeled datasets (Iris, Wine and Breast Cancer); this equivalence is discussed in section 3.2. When dealing with imperfectly labeled data, the same γ_i-EKNN model is also competitive. Letting sources label imperfectly gave better results.

Table 1. Mean accuracy over 100 iterations on perfectly labeled (Iris, Wine, Breast Cancer) and imperfectly labeled (Credal Bird-10 public) datasets.

Dataset	K-NN	γ_q-EKNN	γ-EKNN	γ_i-EKNN
Iris	**0.965** ± 0.006	0.963 ± 0.006	0.964 ± 0.006	0.963 ± 0.006
Wine	**0.737** ± 0.013	0.696 ± 0.012	0.704 ± 0.012	0.696 ± 0.012
Breast Cancer	0.927 ± 0.004	**0.928** ± 0.004	**0.928** ± 0.004	**0.928** ± 0.004
Credal Bird-10 public	0.383 ± 0.015	0.389 ± 0.014	0.411 ± 0.014	**0.412** ± 0.014

4.2 Experiment on noised real world datasets

In this experiment both Iris and Wine datasets have been noised as this is a common point in the literature. A noise parameter $\epsilon = [0, 1]$ is defined and the observations are randomly selected in order to have a proportion of noisy labels equal to ϵ. For each selected observation another singleton is randomly selected and a random mass assigned. The remaining mass is evenly distributed among all other elements. For non-evidential classifiers, the singleton which maximizes the pignistic probability of this new mass is the new label. The Iris and Wine datasets were altered with ϵ equal to 0.5. In this experiment, the mean accuracy of different models is compared using active learning. The models are as follows: K-NN based on a distance weight with 7 nearest neighbors, Logistic Regression with newton-cg for optimization and Random Forest, all used with the scikit-learn default parameters [9]. They are compared to γ_i-EKNN presented in this paper using 7 nearest neighbors. Both experiments used 8 randomly labeled instances (there must be more than K labeled instances) and 20 active learning queries were performed according to uncertainty sampling. The mean accuracy is calculated over 100 iterations.

Figure 1 shows that γ_i-EKNN achieves a mean accuracy of about 0.9 on Iris dataset with only 28 labeled instances, a 30% performance improvement over K-NN due to less significant alteration of the real labels. The distance between the mean accuracy of γ_i-EKNN and K-NN is also greater as the queries number

[1] https://archive.ics.uci.edu

[2] Formula: $[\bar{x} - {}_{1,96}\frac{\mathcal{S}}{\sqrt{n}}; \bar{x} + {}_{1,96}\frac{\mathcal{S}}{\sqrt{n}}]$, with n the size of the sample, $\bar{x}$ its mean and $\mathcal{S}$ the standard deviation of the serie. This formula is used because it is a mean over 100 experiments and not a single proportion.

increases, which means that the model manages to select better instances to label while using the same uncertainty sampling. The same figure 1 shows less optimistic results on the Wine dataset, but still with a dominance of γ_i-EKNN over its non-evidential version. One must be careful with the results, even if the noisy data are distributed in the same way, the labels used for the non-evidential classifiers do not contain the same information as the labels used for γ_i-EKNN, making the comparison more difficult. Apart from the noise, one of the objectives of the paper is to find out whether by adding information during the labeling phase, interesting results can be obtained with a low labeling cost, which leads to the experiment presented in section 4.3.

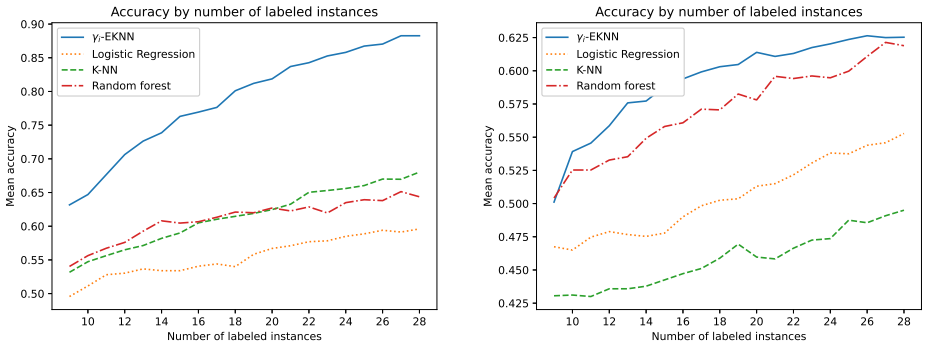


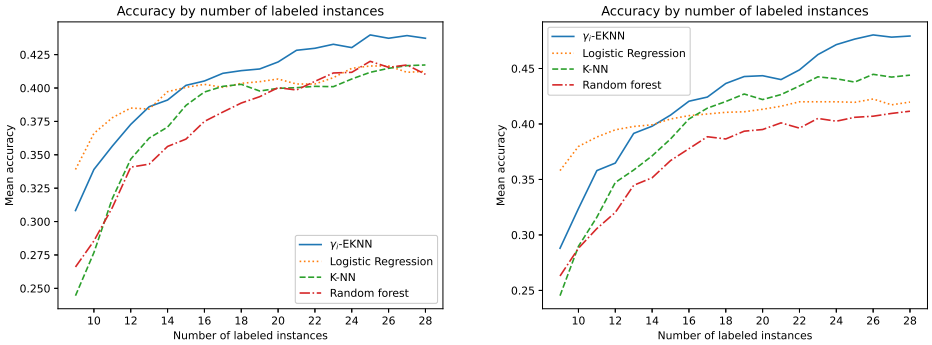
Fig. 1. Mean accuracy by number of labeled instances with 50% noise, on the Iris (a) and Wine (b) datasets.

4.3 Experiment on imperfectly labeled datasets

So far, perfectly labeled datasets have been used for comparison. In this section, a procedure is proposed that can this time be fully compared to non-evidential methods as the labels are unchanged but fundamentally imperfect. To show a real application of the proposed method, we need to train on uncertain and imprecise labels. With the imperfectly labeled dataset of 10 classes introduced in section 3.3 and dimensionally reduced to 512 variables on $\mathcal{X}$, the model is compared to its non-evidential version. The same models and active learning steps as in the experiment 4.2 are used. Differences are present in the datasets, imperfectly labeled by the contributors.

Figure 2 represents the mean accuracy of 100 iterations on both *Credal Bird-10 irisa* and *Credal Bird-10 public* datasets. With 28 labeled instances, EK-NN performs better than its non-evidential version, with around 0.44 accuracy on *Credal Bird-10 irisa* and 0.48 on *Credal Bird-10 public* compared to 0.41 and 0.44 for K-NN respectively. The comparison between the two datasets also shows

that the results may vary greatly depending on the labels, even with the same models. Here, two different populations labeled the same pictures, members of a laboratory and crowdsourcing contributors. This difference produces changes in the results with, in all cases, better results for EK-NN.



(a) (b)

Fig. 2. Mean accuracy by number of labeled instances, on the Credal Bird-10 irisa (a) and Credal Bird-10 public (b) datasets.

5 Conclusion

This study presents a model for efficient learning from a small amount of data derived from imperfect human contributions. It is proposed to couple the theory of belief functions, to model the uncertainty and imprecision of the data, with an active learning algorithm using only a fraction of the labeled data. In particular, our work focuses on the labeling method and how information can be added to allow the learning phase to work more efficiently and at lower cost. A version of the evidential K-nearest neighbors model is proposed, offering a new computation for the parameter γ and allowing to recover an equivalence with the original model in the case of imperfectly labeled data.

To validate this approach, experiments were first conducted on noisy data sets (section 4.2). Very optimistic results are obtained with good performances of the credibility classifiers. However, as the nature of the noise makes it difficult to compare a credibility classifier with its classical version, two new imperfectly labeled datasets on bird species were produced via crowdsourcing to test the model on real data. Few labeled images are used in section 4.3 for decent performance. The quality of the labeling, which depends on the oracle and the model used to represent the imperfection, has a strong influence on the final performance, and can make the results vary more significantly by improving the quality of the labels rather than the quality of the model itself. In future work, we plan to

study how to maximize the quality of the imperfection contained in the labels, by working on its modelisation or on the interface allowing even an inexperienced user to give a relevant uncertain and imprecise answer. Improvement could also be done with active learning, taking into account at the sampling step, that the model can give an evidential answer.

References

1. Dempster, A.P.: Upper and Lower Probabilities Induced by a Multivalued Mapping. The Annals of Mathematical Statistics **38**(2), 325 – 339 (1967)
2. Denoeux, T., Kanjanatarakul, O., Sriboonchitta, S.: A New Evidential K-Nearest Neighbor Rule based on Contextual Discounting with Partially Supervised learning. International Journal of Approximate Reasoning **113**, 287–302 (2019)
3. Denœux, T.: A k-nearest neighbor classification rule based on dempster-shafer theory. Systems, Man and Cybernetics, IEEE Transactions on **219** (1995)
4. Denœux, T., Zouhal, L.: Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets and Systems **122**, 409–424 (09 2001)
5. Dudani, S.A.: The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics **SMC-6**(4), 325–327 (1976)
6. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. Technical Reoort 4. USAF, School of Aviation Medicine, Randolph Field (1951)
7. Fredriksson, T., Mattos, D.I., Bosch, J., Olsson, H.H.: Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In: Morisio, M., Torchiano, M., Jedlitschka, A. (eds.) Product-Focused Software Process Improvement. pp. 202–216. Springer International Publishing, Cham (2020)
8. Jousselme, A.L., Grenier, D., Éloi Bossé: A new distance between two bodies of evidence. Information Fusion **2**(2), 91–101 (2001)
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
10. Roh, Y., Heo, G., Whang, S.E.: A survey on data collection for machine learning: A big data - ai integration perspective. IEEE Transactions on Knowledge and Data Engineering **33**(4), 1328–1347 (2021)
11. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
12. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
13. Thierry, C., Hoarau, A., Martin, A., Dubois, J.C., Le Gall, Y.: Real bird dataset with imprecise and uncertain values. In: 7th International Conference on Belief Functions (2022)
14. Zhu, D., Martin, A., Le Gall, Y., Dubois, J.C., Lemaire, V.: Evidential Nearest Neighbours in Active Learning. In: Worksop on Interactive Adaptive Learning (IAL) - ECML-PKDD. Bilbao, Spain (Sep 2021)