Uncertain data in learning: challenges and opportunities

Sébastien Destercke SEBASTIEN.DESTERCKE@HDS.UTC.FR
HEUDIASYC - UMR CNRS 7253, Université de Technologie de Compiègne, 57 avenue de Landshut, 60203 COMPIEGNE CEDEX - FRANCE

Editor: Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo and Lars Carlsson

Abstract

Dealing with uncertain data in statistical estimation problems or in machine learning is not really a new issue. However, such uncertainty has so far mostly been modelled either as sets, being called for instance coarse data or partial labels, or as probability distributions over data values, being called for instance soft labels. Integrating this uncertainty in the learning process can be challenging, but also rewarding, as it can improve both the quality of the made predictions as well as our understanding of the obtained model. Within this setting, rich uncertainty models generalizing both probabilities and sets offer both new challenges and opportunities, and I will summarise some of them in this short note.

Keywords: Uncertain data, imprecision, credal representation, learning

1. Introduction

There are various reasons to get interested in how to model data uncertainty and reason with it. Data uncertainty may indeed appear in various situations. For instance, measurement tools may only provide measurements up to a given precision, or expert labelling some data may be uncertain in front of ambiguous or previously unseen situations. In other settings, it is quite common to only ask for part of the information: for example, in preference learning, users will typically only give preferences over a subset of alternatives rather than over all of them. Finally, in co-learning or self-supervised learning settings, unlabelled data are typically labelled according to some model predictions, which are themselves uncertain.

Dealing with such uncertain data in statistics and machine learning is of course not a new issue. Formally, we will be uncertain about the value that a data can take over a space $\mathcal{Y}$ of possible values (typically a finite set in multi-class problems or a subset of $\mathbb{R}$ in a regression problem). For most works concerning this issue, how this uncertain data is modelled and accounted for can be divided into two main options:

- The first considers that a piece of uncertain data is modelled by a set $E \subseteq \mathcal{Y}$ of possible values, one of which is the true one. This corresponds, for instance, to the coarse data described by Rubin and Little seminal work (Little and Rubin, 2019). Another set-valued problem explored by several authors in supervised machine learning is to deal with partial class labels (Cour et al., 2011; Liu and Dietterich, 2014).

- The second considers that an uncertain data can be modelled as a probability distribution p over the possible values $\mathcal{Y}$ a data can take. This is a usual assumption in database literature (Aggarwal and Philip, 2008), and such a modelling is also commonly used in machine learning to build pseudo-labels (a.k.a. soft labels) whose use often results in more regular, better calibrated models (Müller et al., 2019).

While probabilities and sets are often considered apart from each other, uncertainty theories such as belief functions (Shafer, 1976) and imprecise probabilities (Augustin et al., 2014) generalize both by considering more expressive uncertainty theories. They therefore offer nice theoretical frameworks to build a unified theory of learning with uncertain data. Moreover, many recent works (Martin, 2019; Grünwald, 2018; Vovk and Petej, 2014) point out that providing generic calibrated predictions in a number of situations actually requires to use more expressive models than probability, or at least to provide imprecise statements about those probabilities.

In this short note, I will first discuss the nature and modelling of data uncertainty, and in particular why convex sets of probabilities, a.k.a. credal sets, are interesting models to consider. This will be done in Section 2. Focusing on the case of purely imprecise, set-valued data, I will then discuss in Section 3 the challenges, both in terms of computation and interpretation, that may arise from using such uncertainty models. To finish on a more optimistic note, I will discuss in Section 4 how using rich uncertainty models can actually be beneficial to the learning process, and present interesting opportunities for the future.

2. Discussing data uncertainty and the way to model it

When thinking about modelling the uncertainty we have about a quantity, people have considered quite a number of distinctions: examples of known distinctions include, for instance

- aleatoric versus epistemic uncertainty, that is uncertainty due to te intrinsic randomness of a process versus uncertainty due to lack of knowledge;
- objective versus subjective uncertainty, that is whether uncertainty can be measured as an objective quantity, or only depends on the agent that tries to model it;
- generic versus singular uncertainty, that is whether the uncertainty concerns a whole population or a repeatable experiment, or a unique non-repeatable situation;
- reducible versus non-reducible uncertainty, that is whether uncertainty can be reduced or even suppressed by collecting more information, or if it will remain whatever the amount of information we will obtain in the future.

Such distinctions are useful to convey the idea that uncertainty is multi-faceted, but can be the topic of endless debates about their exact meaning and nature, and it is not my goal here to enter into such kind of debates.

However, when concerned with data, or rather datum uncertainty, it is fair to say that our uncertainty usually[1] concerns a fixed but ill-known value Y, in the sense that if we had prefect, trustworthy information about the piece of data, we would end up with a single precise value. This uncertainty may be reducible or not, depending on whether we can acquire more measurements, but the quantity we are interested in is mostly of a non-statistical, singular nature.

Among other things, this means that there is no reason, *a priori*, to prefer a probabilistic model when it comes to the problem of describing our uncertainty about Y mathematically. While using probabilities for non-random quantities and non-statistical settings has been justified by De Finetti (De Finetti, 2017) and others, there are good arguments indicating that one may prefer more general models (Walley, 1991).

[1] even this can be questioned, and I will briefly discuss it in Section 3

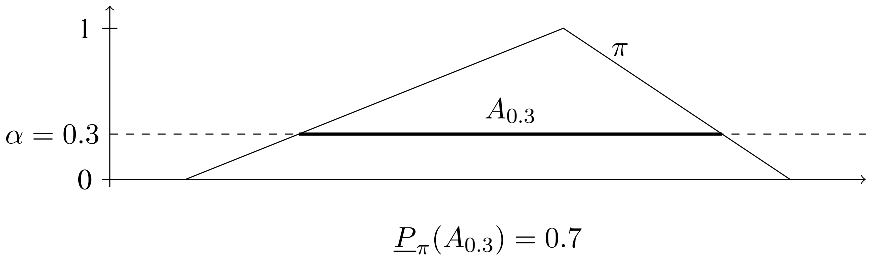


Figure 1: Illustration of possibility distributions on $\mathbb{R}$

Imprecise probabilities are such general models, that enrich probabilities by considering as their basic uncertainty models (convex) sets of probabilities, a.k.a., credal sets, usually denoted as $\mathcal{P}$. If $\Delta_{\mathcal{Y}}$ denotes the set of all possible probabilities over $\mathcal{Y}$, then a credal set $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ is simply a subset of $\Delta_{\mathcal{Y}}$. Given a credal set $\mathcal{P}$, the lower and upper probabilities of an event $A \subseteq \mathcal{Y}$ are simply defined as

$$\overline{P}(A) = \sup_{P \in \mathcal{P}} P(A), \quad \underline{P}(A) = \inf_{P \in \mathcal{P}} P(A). \tag{1}$$

They are dual, in the sense that $\overline{P}(A) = 1 - \underline{P}(A)$ Such models include both probabilities (in which case $\overline{P} = \underline{P}$) and sets (in which case the set E corresponds to $\underline{P}(A) = 1$ for any $E \subseteq A$, zero otherwise). Conversely, given an upper measure $\overline{P}$, one can consider the associated credal set

$$\mathcal{P}_{\overline{P}} = \{P : P(A) \leq \overline{P}(A)\}$$

of dominated probabilities, and likewise for lower probabilities, considering the set of dominating probabilities[2].

For practical purposes, one will often consider particular credal sets, whose structure offers some advantages or interests in terms of interpretation, computations and mathematical properties (Destercke and Dubois, 2014). I will now review some of those models, focusing on those that are closely linked to conformal prediction and Venn-Abers predictors (Johansson et al., 2019).

2.1. Possibility distributions

A possibility distribution (Dubois and Prade, 1992) is a positive mapping $\pi : \mathcal{Y} \to [0, 1]$ such that its maximum is one, i.e., $\max_{x \in \mathcal{X}} \pi(x) = 1$. From such a distribution is then defined a maxitive or supremum-preserving upper probability such that

$$\overline{P}_{\pi}(A) = \sup_{x \in A} \pi(x) \tag{2}$$

A particularly interesting feature of the credal set $\mathcal{P}_{\pi}$ induced by a possibility distribution is that it can be totally characterised by lower confidence values provided over sequences of nested intervals. In particular, if we define the cut A_{α} as $\{x : \pi(x) \geq \alpha\}$, the set $\mathcal{P}_{\pi}$ can be defined as

$$\mathcal{P}_{\pi} = \{P : \forall \alpha \in [0, 1], \underline{P}_{\pi}(A_{\alpha}) \geq 1 - \alpha\}.$$

[2] In general the upper/lower measure inducing a credal set do not need to be their lower/upper envelope, i.e., using Equation (1) on $\mathcal{P}_{\overline{P}}$ may not give back $\overline{P}$, but it will be the case for all models considered here

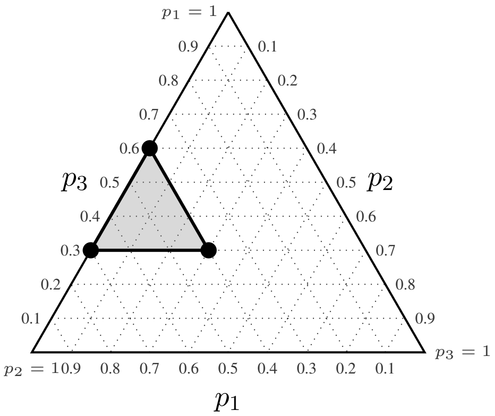


Figure 2: Probability set induced by imprecise probabilistic assignment.

There is a strong similarity between the sets $\{x : \pi(x) \geq \alpha\}$ and the p-values used in conformal predictions. In fact, one can show that the p-values derived by conformal predictors can be interpreted as possibility degrees (Lienen et al., 2022; Cella and Martin, 2022). This means that conformal procedures actually produce calibrated model of data uncertainty in the form of possibility distributions, that can in turn be used in learning procedures. Such uses will be mentioned and briefly discussed later on.

It should be noticed, however, that possibility distributions are not able to model precise probabilities, as the interval $[\underline{P}_\pi, \overline{P}_\pi]$ will always be of the kind $[0, \alpha]$ or $[\beta, 1]$, therefore not allowing to model the precise interval $[\gamma, \gamma]$ whenever $\gamma \in (0, 1)$.

2.2. Imprecise probability assignments (IPA)

When $\mathcal{Y}$ is a discrete space, another commonly used credal representation is the one of imprecise probability assignments (De Campos et al., 1994), that consists in specifying probabilistic bounds $[\underline{p}(y), \overline{p}(y)]$ over each singleton of the space $\mathcal{Y}$. The lower and upper probabilities obtained from such models are

$$\underline{P}(A) = \max\{\sum_{y \in A} \underline{p}(y), 1 - \sum_{y \notin A} \overline{p}(y)\}, \quad \overline{P}(A) = \min\{\sum_{y \in A} \overline{p}(y), 1 - \sum_{y \notin A} \underline{p}(y)\},$$

from which can be derived a corresponding credal set $\mathcal{P}_{[\underline{p},\overline{p}]}$. Figure 2 provides an illustration in barycentric coordinates where $p_i = p(y_i)$ and where $p(y_1) \in [0.3, 0.6]$, $p(y_2) \in [0.4, 0.7]$, $p(y_3) \in [0, 0.3]$. In contrast with credal sets induced by possibility distributions, credal sets induced by imprecise probability assignments include both sets E ($[\underline{p}(y), \overline{p}(y)] = [0, 1]$ for any $y \in E$) and probabilities p ($[\underline{p}(y), \overline{p}(y)] = [p(y), p(y)]$).

Imprecise probability assignments seem to be ideal candidates to model the predictive uncertainty derived from the use of Venn-Abers predictors (Vovk and Petej, 2014), as those typically provide probability bounds for each possible alternatives. Moreover, while being slightly more complex representations than possibility distributions, imprecise probability assignments still enjoy many nice mathematical properties, meaning that the computational cost of using them in learning

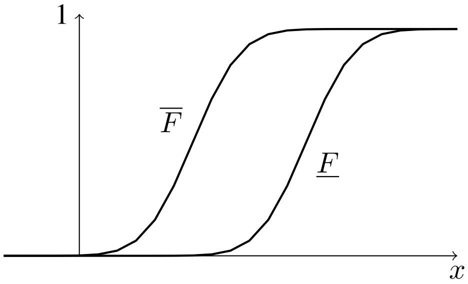


Figure 3: Illustration of p-box.

procedures may be limited. However, to my knowledge, the line of research consisting of using such models within learning procedures has not been investigated so far.

2.3. Imprecise cumulative distributions (p-boxes)

The final model I will mention is the one commonly known under the name p-box (or probability box), that considers sets of probabilities whose cumulative distributions are included between two bounding cumulative ones. More precisely, if $\mathcal{Y} = \mathbb{R}$, a p-box on $\mathcal{Y}$ is a pair $[\underline{F}, \overline{F}]$ of increasing functions such that $\underline{F}(y) \leq \overline{F}(y)$ and $\underline{F}(\infty) = \overline{F}(\infty) = 1$. The induced credal set is then defined as

$$\mathcal{P}_{[\underline{F},\overline{F}]} = \{P : \underline{F}(y) \leq F_P(y) \leq \overline{F}(y)\}.$$

Figure 3 provides a picture of a p-box. It is clear that p-boxes are reminiscent of the predictive distributions proposed for example in the conformal setting by Vovk et al. (2018). They include both standard probabilities (when $\underline{F} = \overline{F}$) and intervals $E = [a, b]$ (modelled by $\overline{F}(y) = 1$ for any $y \geq a$, zero else, and $\underline{F}(y) = 1$ for any $y \geq b$, zero else). In order for them to model arbitrary sets, one needs to extend their definition to any ordering over the space $\mathcal{Y}$ (Destercke et al., 2008).

3. Learning from credal data: challenges

I will now briefly comment on how uncertain data fits into the learning setting, and some of the challenges associated to using richer uncertainty models. I will mainly consider a loss minimisation perspective, even if the same kind of questions can be considered in statistical estimation (Couso and Dubois, 2018).

Set-valued case

We are now concerned with learning a model $h_\theta : \mathcal{X} \to \mathcal{Y}$ from N observations $\mathcal{D} = \left\{ (\boldsymbol{x}_n, y_n) \right\}_{n=1}^{N}$. When those observations are precise, and assuming that we quantify the loss of predicting through a loss function $\ell : \mathcal{Y} \to \mathcal{Y}$ such that $\ell(y, h_\theta(x))$ is the loss of predicting $h_\theta(x)$ when observing y, an optimal model is usually obtained by minimizing the empirical risk

$$\mathcal{R}_{emp}(\theta) := \frac{1}{N} \sum_{n=1}^{N} \ell\big(y_n, h_\theta(\boldsymbol{x}_n)\big).$$

If our knowledge of y is replaced by an interval or a set E, this equation is no longer well defined, as the loss becomes itself ill-defined and can take various values. Common replacements used to obtain again a precisely defined loss include

- Optimistic (Minimin) approach (Hüllermeier, 2014; Cour et al., 2011):

$$\ell_{opt}(E, h_\theta(\boldsymbol{x})) = \min\{\ell(y, h_\theta(\boldsymbol{x}))|y \in E\}$$

- Pessimistic (Minimax) approach (Guillaume et al., 2017):

$$\ell_{pes}(E, h_\theta(\boldsymbol{x})) = \max\{\ell(y, h_\theta(\boldsymbol{x}))|y \in E\}$$

- ”EM-like” or averaging/weighting approaches[3]

$$\ell_w(E, h_\theta(\boldsymbol{x})) = \sum_{y \in E} w_y \ell(y, h_\theta(\boldsymbol{x})).$$

The optimistic approach will pick the data and the model that are the most favourable to us. In a sense, they assume that the true value of imprecise and uncertain data are distributed favourably with respect to our hypothesis. In contrast, the pessimistic approach will try to find the model that behaves as well as possible in all possible scenarios or for all possible true values of the uncertain data. Such a choice may cover, for instance, cases where we want to be robust against data uncertainty. We could think for instance of training performed under some nominal conditions x_i, but where one wants the model to remain as efficient as possible when conditions vary within an interval $x_i \pm \epsilon$.

As an illustration, considers the case of binary classification $\mathcal{Y} \in \{0, 1\}$ with a log-loss function

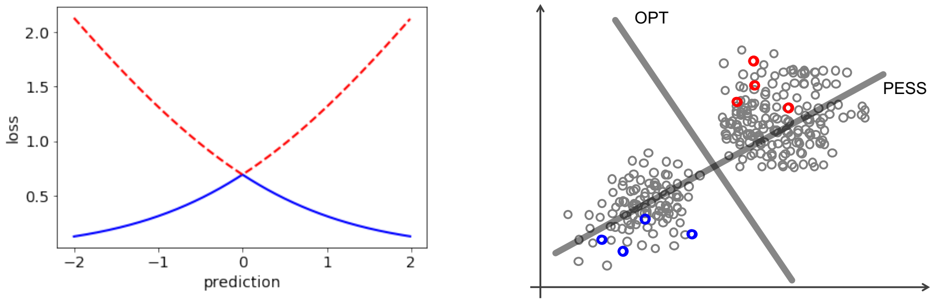
$$\ell(y, p) = -\log\big(py + (1-p)(1-y)\big) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{if } y = 0, \end{cases}$$

where $h_\theta(x) = p$ is the predicted probability $p(1)$. Figure 4 shows the behaviour of the loss function and of the obtained models for the optimistic and pessimistic versions. One can clearly see that the choice of the precise loss function induced by an imprecise observation can have a huge impact on the end-result (in this extreme case, the two models are orthogonal). Beyond the obvious computational problems that may pose the optimisation of the various loss functions, another challenge is to explore the underlying statistical hypothesis that correspond to each of the possible choices, as well as to quantify the impact that such choices could have on the end-result. Hüllermeier et al. (2019) provide some discussions along this line.

Credal-valued case

In the credal case, our data uncertainty can sometimes be represented by a set E, but would in general be represented by a credal set $\mathcal{P}$, possibly induced by one of the models described in Section 2. In such a case, the same strategies as the ones we just described can still be applied, in particular if

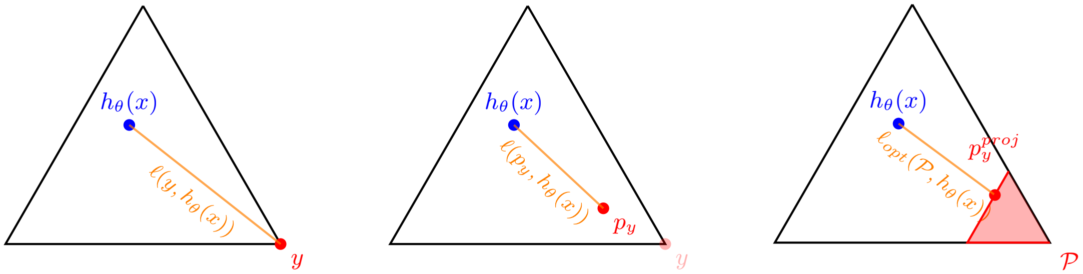
[3] With likelihood $\sim L_{av}(\theta|(x, E)) = P((x, E)|\theta)$ (Denoeux, 2013)



(a) optimistic (blue) and pessmistic (red) log-loss

(b) obtained model

Figure 4: Set-valued data and logistic regression



(a) Hard label

(b) Soft label

(c) Credal labels

Figure 5: Extensions of loss functions

we consider a loss function $\ell : \Delta_\mathcal{Y} \to \Delta_\mathcal{Y}$, and if our model $h_\theta(x)$ returns a probability over $\mathcal{Y}$. In such a situation we can easily define, at least theoretically, the extended losses

$$\ell_{opt}(\mathcal{P}, h_\theta(\boldsymbol{x})) = \min\{\ell(p, h_\theta(\boldsymbol{x}))|p \in \mathcal{P}\},$$

$$\ell_{pes}(\mathcal{P}, h_\theta(\boldsymbol{x})) = \max\{\ell(p, h_\theta(\boldsymbol{x}))|p \in \mathcal{P}\}.$$

Note that such requirements will be often met in practice, for instance when one uses cross-entropy as a loss function. Figure 5 illustrates what becomes of the loss function in the various cases. While using a soft/probabilistic labels will tend to reduce the loss incurred by $h_\theta(x)$, and therefore the subsequent correction of the model, this loss will always be positive. In contrast, the optimistic loss may actually reduce to zero (and therefore to no model correction after having observed $\mathcal{Y}$) if $h_\theta(x) \in \mathcal{P}$. This is especially interesting if $\mathcal{P}$ is well-calibrated, as it is when resulting from a conformal procedure. In theory, the case of credal labels is therefore not very different from the set-valued case, but may present additional computational challenges. Lienen and Hüllermeier (2021) however show that, in the case of possibilistic models, this is quite doable. Note that while we have here mostly focused on uncertainty in output variables, the ideas presented here readily extend to uncertain inputs (Hüllermeier et al., 2019), at least in theory.

A brief discussion about uncertain data in conformal approaches

The previous sections have described how the estimation of a predictive model could be achieved in spite of having uncertain data. Considering uncertain data within a (inductive) conformal setting however raises at least an additional question, which is how to deal with uncertain data in the calibration data set $\mathcal{D}_{cal}$.

Formally speaking, if a precise data (x_i, y_i) in the calibration set is mapped to a precise conformity score $\alpha_i \in \mathbb{R}$, then an uncertain data $(x_i, \mathcal{P}_i)$ would be mapped to a corresponding "credal conformity score" $\mathcal{A}_i$. In particular, if (x_i, y_i) is set-valued, the conformity score would be a set $A_i \subseteq \mathbb{R}$ having a minimal $\underline{\alpha}_i = \min A_i$ and a maximal $\overline{\alpha}_i = \min A_i$ conformity score. Options to obtain a final predictions could then be for example to

- systematically select $\alpha_i = \underline{\alpha}_i$, so as to maximize the size of the produced conformal prediction, but at the expense of being valid only in a conservative way;
- provide inner (by considering $\alpha_i = \overline{\alpha}_i$) and outer approximation of what would be the actual but ill-known conformal prediction;
- to find an adequate strategy to still ensure validity of the produced prediction, probably with the need to specify some hypothesis about how the set-valued data have been produced.

While such a situation is less likely to happen for standard multi-class and regression problems (although hard to label cases and imprecise/noisy output observations are not uncommon in those problems), they are much more likely to happen in more complex issues, such as frameworks involving complex predictions (e.g., multi-task problems, ranking problems) or frameworks involving repeated use of data such as in stacking methods.

4. Learning from uncertain data: opportunities

So far, I have mainly mentioned some issues regarding the integration of uncertain data to the learning process, such as how to find an adequate adaptation of loss functions, or how to compute with such adaptations.

However, accurately modelling data uncertainty and integrating it in the learning process can actually be quite beneficial. An obvious advantage is that it would be helpful to know to which extent a data point should impact our learning process: for instance, if we come back to Figure 5, it is clear that the bigger will be $Q \subseteq \Delta_{\mathcal{Y}}$, i.e., the weaker will be the information we have about a data, the less change it will induce on our final model. In particular, if our uncertainty about a data point boils down to ignorance, that is if $Q = \Delta_{\mathcal{Y}}$, no change at all will be performed, at least if we pick $\ell_{opt}(E, h_\theta(\boldsymbol{x}))$.

Immediate techniques where this applies are those where one starts with some labelled $\mathcal{D}_L(x_i, y_i)$, $i = 1, \ldots, n$ and a number of m observed unlabelled data $\mathcal{D}_U = (x_j, \cdot)$, $j = n+1, \ldots, n+m$, and use a predictive model to label some data in $\mathcal{D}_U$. Typically, some selected items (e.g. using uncertainty quantification tools) are removed from $\mathcal{D}_U$ and integrated into $\mathcal{D}_L$ by assigning them hard labels. However, an alternative would be to use (calibrated) credal uncertainty models to obtain a data set $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, \mathcal{P}_{n+1}, \ldots, \mathcal{P}_{n+m})$ to which can then be applied to techniques of Section 3. Figure 6 illustrates this idea, where pseudo-labelled samples which are associated to credal predictions are just kept in the loop. Such strategies have already been proven efficient in the

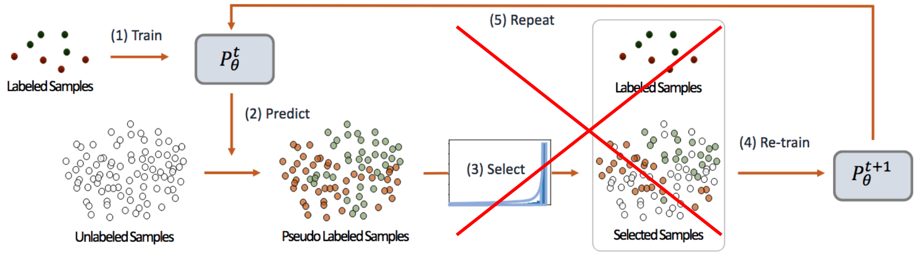


Figure 6: Self-supervised learning with credal approaches (picture from Cascante-Bonilla et al. (2021))

case of possibility distributions (Lienen and Hüllermeier, 2021), and even more as those distributions are issued from conformal predictors (Lienen et al., 2022).

Those researches show that uncertainty quantification methods, beyond satisfying the necessary task of quantifying uncertainty, can actually be seen as opportunities to improve learned models predictive capabilities. This opens up multiple research paths, such as providing extensions to other settings (for instance by considering interval bounds derived by Venn-Abers predictors).

References

Charu C Aggarwal and S Yu Philip. A survey of uncertain data algorithms and applications. *IEEE Transactions on knowledge and data engineering*, 21(5):609–623, 2008.

Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920, 2021.

Leonardo Cella and Ryan Martin. Validity, consonant plausibility measures, and conformal prediction. *International Journal of Approximate Reasoning*, 141:110–130, 2022.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Inés Couso and Didier Dubois. A general framework for maximizing likelihood under incomplete data. *International Journal of Approximate Reasoning*, 93:238–260, 2018.

Luis M De Campos, Juan F Huete, and Serafin Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2 (02):167–196, 1994.

Bruno De Finetti. *Theory of probability: A critical introductory treatment*, volume 6. John Wiley & Sons, 2017.

Thierry Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on knowledge and data engineering*, 25(1):119–130, 2013.

Sébastien Destercke and Didier Dubois. Special cases. *Introduction to Imprecise Probabilities*, (chapter 4):79–91, 2014.

Sébastien Destercke, Didier Dubois, and Eric Chojnacki. Unifying practical uncertainty representations–i: Generalized p-boxes. *International Journal of Approximate Reasoning*, 49 (3):649–663, 2008.

Didier Dubois and Henri Prade. When upper probabilities are possibility measures. *Fuzzy sets and systems*, 49(1):65–74, 1992.

Peter Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 195:47–63, 2018.

Romain Guillaume, Inés Couso, and Didier Dubois. Maximum likelihood with coarse data based on robust optimisation. In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pages 169–180, 2017.

Eyke Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

Eyke Hüllermeier, Sébastien Destercke, and Ines Couso. Learning from imprecise data: adjustments of optimistic and pessimistic variants. In *International Conference on Scalable Uncertainty Management*, pages 266–279. Springer, 2019.

Ulf Johansson, Tuwe Löfström, and Henrik Boström. Calibrating probability estimation trees using venn-abers predictors. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 28–36. SIAM, 2019.

Julian Lienen and Eyke Hüllermeier. Credal self-supervised learning. *Advances in Neural Information Processing Systems*, 34:14370–14382, 2021.

Julian Lienen, Caglar Demir, and Eyke Hüllermeier. Conformal credal self-supervised learning. *arXiv preprint arXiv:2205.15239*, 2022.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Liping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *International Conference on Machine Learning*, pages 1629–1637. PMLR, 2014.

Ryan Martin. False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113:39–73, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 829–838, 2014.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 37–51. PMLR, 2018.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.