Méthode crédibiliste pour l'extraction d'incertitudes sans dépendance aux observations

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

*University of Rennes, CNRS, IRISA, DRUID, France
https://www.irisa.fr/
**Orange Labs, Lannion, France

Résumé. Les recherches récentes en apprentissage actif, et plus précisément en échantillonnage par incertitude, se sont concentrées sur la décomposition de l'incertitude du modèle en incertitudes réductibles et irréductibles. Dans cet article, nous proposons de simplifier la phase de calcul et de supprimer la dépendance vis-à-vis des observations, mais surtout de prendre en compte l'incertitude déjà présente dans les étiquettes, *i.e.* l'incertitude des oracles. La stratégie proposée, l'échantillonnage par incertitude de Klir, traite également le dilemme d'exploration-exploitation, en utilisant la théorie des fonctions de croyance.

1 Introduction

Pour des raisons d'efficacité, de coût ou de réduction d'énergie en apprentissage automatique ou en apprentissage profond, l'une des questions importantes est liée à la quantité de données étiquetées. L'apprentissage actif (Settles, 2009) est une partie de l'apprentissage automatique dans laquelle l'apprenant peut choisir les observations à étiqueter afin de ne travailler qu'avec une fraction du jeu de données étiquetées. Parmi toutes les stratégies proposées dans la littérature, décrites par (Settles, 2009) et (Aggarwal et al., 2014), l'une des plus connues est l'échantillonnage par incertitude (Nguyen et al., 2022). La plupart des mesures utilisées pour quantifier cette incertitude, comme l'entropie, sont jusqu'à présent probabilistes. Dans cet article, nous proposons d'utiliser un cadre plus large de modélisation de l'incertitude qui généralise les probabilités.

Comme le proposent les articles récents de (Hüllermeier et Waegeman, 2021; Kendall et Gal, 2017; Senge et al., 2014), l'incertitude peut être décomposée en deux notions distinctes : l'incertitude épistémique et l'incertitude aléatoire. L'incertitude aléatoire découle de la propriété stochastique de l'événement et n'est donc pas réductible, tandis que l'incertitude épistémique est liée à un manque de connaissances et peut être réduite. Les calculs proposés dépendent de la prédiction du modèle, mais aussi des observations. Nous proposons dans cet article de supprimer la dépendance directe vis-à-vis des observations et de n'utiliser que la sortie du modèle pour obtenir des résultats similaires. Cette représentation aborde également la question de l'exploration-exploitation en apprentissage actif, avec la possibilité de choisir l'un ou l'autre, ou même un compromis comme (Bondu et al., 2010).

Le processus d'étiquetage est souvent réalisé par des humains sans faire de différence entre une étiquette donnée par quelqu'un qui a hésité pendant longtemps et une étiquette donnée par quelqu'un qui n'a aucun doute, et donc l'incertitude peut déjà exister dans les étiquettes. Dans le cas de la classification supervisée, plusieurs modèles sont maintenant capables de traiter ces étiquettes incertaines (Denoeux, 1995; Hoarau et al., 2023; Denoeux et Bjanger, 2000). L'objectif principal de l'article, en plus de ne pas dépendre des observations et d'aborder la question de l'exploration-exploitation, est de prendre en compte dans en échantillonnage l'incertitude déjà présente dans les étiquettes et induite par les sources qui étiquettent les observations (*i.e.* "Oracles" en apprentissage actif).

Nous proposons dans cette étude une stratégie d'échantillonnage capable de représenter une décomposition des incertitudes du modèle par rapport à l'incertitude déjà présente dans les étiquettes. Cette incertitude est construite à partir de deux incertitudes différentes, la discorde qui est le degré d'autoconfusion des informations et la non-spécificité qui est le degré d'ignorance des informations, dans les résultats du modèle.

Le document est organisé comme suit ; la section 2 introduit quelques notions importantes d'étiquetage imparfait et la modélisation de ces étiquettes plus riches à l'aide de la théorie des fonctions de croyance. L'approche habituelle de l'échantillonnage par incertitude (Nguyen et al., 2022) est également rappelée et la section 3 décrit la séparation entre les incertitudes aléatoires et épistémiques. La section 4 présente la stratégie proposée, puis la section 5 discute et conclut l'article. Les expériences réalisées dans cet article[1] sont présentées en annexe, afin d'éviter de longues explications, puisque l'objectif de l'article ne réside pas dans cette partie. En outre, les incertitudes sont cartographiées sur des représentations 2D, mais l'objectif est de servir l'apprentissage actif plus tard.

2 Préliminaires

2.1 Étiquetage imparfait

La plupart des jeux de données utilisés pour la classification considèrent des étiquettes dures, avec une appartenance binaire où l'observation est soit membre de la classe, soit non membre. Dans le présent document, nous qualifions d'étiquettes riches les éléments de réponse fournis par une source qui peut comporter plusieurs degrés d'imprécision (*i.e.* "*Ça pourrait être un chat*", "*Je ne sais pas*" ou "*J'hésite entre chien et chat, avec une légère préférence pour le chat*)". De tels jeux de données, offrant une incertitude déjà présente dans les étiquettes, existent, mais ne sont pas nombreux. Ces étiquettes sont appelées "riches" dans le présent document, car elles fournissent plus d'informations que les étiquettes "dures" et peuvent être modélisées à l'aide de la théorie des fonctions de croyance.

2.2 Théorie des fonctions de croyance

La théorie des fonctions de croyance (Dempster, 1967; Shafer, 1976), est utilisée dans cette étude pour modéliser l'incertitude et l'imprécision pour les phases d'étiquetage et de prédiction. Soit $\Omega = \{\omega_1, \ldots, \omega_M\}$ le cadre de discernement pour M hypothèses exclusives et exhaustives. On suppose qu'un seul élément de Ω est vrai (hypothèse du monde fermé

[1] `https://github.com/ArthurHoa/evidential-uncertainty-sampling`

rappelée par (Smets et Kennes, 1994)). L'ensemble de puissance 2^Ω est l'ensemble des parties de Ω. Une fonction de masse attribue la croyance qu'une source peut avoir sur les éléments de l'ensemble des parties de Ω, de telle sorte que la somme de toutes les masses soit égale à 1.

$$m : 2^\Omega \rightarrow [0,1], \sum_{A \in 2^\Omega} m(A) = 1. \quad (1)$$

Tout sous-ensemble $A \in 2^\Omega$ tel que $m(A) > 0$ est appelé un *élément focal* de m. L'incertitude est donc représentée par une masse $m(A) < 1$ sur un élément focal A et l'imprécision est représentée par une masse non nulle $m(A) > 0$ sur un élément focal A tel que $|A| > 1$.

Au niveau décisionnel, la probabilité pignistique $BetP$ de (Smets et Kennes, 1994) aide à la prise de décision sur les singletons :

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

Il est également possible de combiner plusieurs fonctions de masse (croyances provenant de différentes sources) en une seule fonction de masse. Si les étiquettes et donc les masses ne sont pas indépendantes, une simple moyenne des fonctions de masse m_j dérivées de N sources peut être définie comme suit :

$$m(A) = \frac{1}{N} \sum_{j=1}^{N} m_j(A), \quad A \in 2^\Omega. \quad (3)$$

• **Exemple 1:** Soit $\Omega = \{Chien, Chat\}$ un cadre de discernement. Une observation étiquetée "Chat" par une source peut être modélisée dans le cadre des fonctions de croyance par la fonction de masse m_1 telle que : $m_1(\{Cat\}) = 1$ et $m_1(A) = 0,\ \forall A \in 2^\Omega \backslash \{Chat\}$.

• **Exemple 2:** Une observation étiquetée "Chat ou Chien" par une source peut être modélisée par la fonction de masse m_2 telle que :$m_2(\{Chat, Chien\}) = 1$ et $m_2(A) = 0$, $\forall A \in 2^\Omega \backslash \{Chat, Chien\}$.

• **Exemple 3:** La fonction de masse moyenne $\bar{m}$ de m_1 et m_2 est : $\bar{m}(\{Chat\}) = 0,5$, $\bar{m}(\{Chat, Chien\}) = 0,5$ et $\bar{m}(A) = 0$ pour tous les autres sous-ensembles A dans 2^Ω. Sa probabilité pignistique $BetP$, utilisée pour la prise de décision, est : $BetP(\{chat\}) = 0,75$ et $BetP(\{Chien\}) = 0,25$.

2.3 Échantillonnage par incertitude

L'apprentissage actif construit itérativement un jeu d'entraînement en sélectionnant les meilleures instances à étiqueter. Le principe est, pour des performances ou un budget donné, d'étiqueter le moins d'observations possible. Parmi toutes les stratégies proposées dans la littérature (Settles, 2009), la méthode la plus connue est l'échantillonnage par incertitude, où la fonction qui définit les instances à étiqueter maximise l'incertitude liée à la prédiction du modèle, comme décrit ci-dessous. Soit $\mathcal{U}$ l'incertitude liée à l'étiquetage d'une nouvelle observation x pour un modèle donné et $\Omega = \{\omega_1, \ldots, \omega_M\}$ l'ensemble des M classes possibles. L'incertitude $\mathcal{U}$ peut être calculée de plusieurs manières, une approche classique consiste à utiliser l'entropie de Shannon :

$$\mathcal{U}(x) = -\sum_{\omega \in \Omega} p(\omega|x) \log[p(\omega|x)], \quad (4)$$

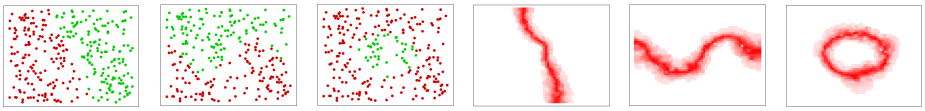


Figure 1: Trois jeux de données à deux classes avec les zones d'incertitude du modèle.

avec $p(\omega|x)$ la probabilité pour x d'appartenir à la classe ω, donnée par le modèle. D'autres critères d'incertitude existent, il est courant d'utiliser la mesure de moindre confiance :

$$\mathcal{U}(x) = 1 - \max_{\omega \in \Omega}[p(\omega|x)]. \tag{5}$$

La figure 1 représente trois jeux de données, dont les classes sont parfaitement séparées. Étant donné un modèle et l'un des critères d'incertitude, nous pouvons calculer l'incertitude de n'importe quel point dans l'espace. Pour chaque jeu de données, les zones d'incertitude du modèle sont représentées, avec plus de rouge pour plus d'incertitude. Il est remarquable que ces zones d'incertitude puissent être comparées à la frontière de décision du modèle.

L'échantillonnage par incertitude consiste à choisir l'observation pour laquelle le modèle est le moins sûr de sa prédiction. C'est l'une des bases de l'apprentissage actif, mais d'autres méthodes permettent d'extraire plus d'informations sur cette incertitude, ce qui conduit à la décomposer en incertitudes épistémiques et aléatoires.

3 Intérêts et limites des incertitudes épistémiques et aléatoires pour l'apprentissage actif

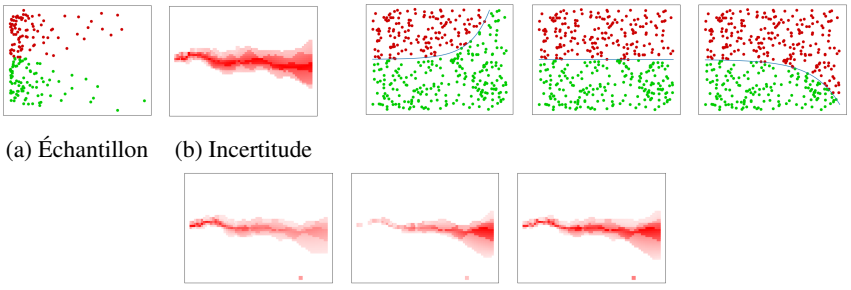
Cette section introduit des éléments supplémentaires pour décomposer l'incertitude du modèle afin qu'il puisse se concentrer sur les observations induisant un gain rapide en performance.

L'incertitude $\mathcal{U}(x)$ peut être séparée en deux incertitudes (Hora, 1996), l'une réductible et l'autre irréductible. Par exemple, le résultat d'un pile ou face est incertain et il n'est pas possible de générer davantage de connaissances pour prédire que la pièce tombera sur pile ou face, cette ignorance est appelée incertitude aléatoire. Le mot *Klaava* en finnois est soit "pile" soit "face", il s'agit d'une incertitude qui peut être résolue en apprenant cette langue, c'est ce qu'on appelle l'incertitude épistémique.

La possibilité de modéliser ces deux incertitudes peut aider à délimiter les zones où il est plus intéressant de fournir des connaissances et où ce n'est pas le cas. L'incertitude totale est souvent représentée comme la somme de l'incertitude épistémique $\mathcal{U}_e(x)$ et de l'incertitude aléatoire $\mathcal{U}_a(x)$: $\mathcal{U}(x) = \mathcal{U}_e(x) + \mathcal{U}_a(x)$.

Pour un problème à deux classes $\Omega = \{0, 1\}$, il est proposé par (Senge et al., 2014) de modéliser l'incertitude, sous le formalisme de (Nguyen et al., 2022), en calculant la plausibilité π d'appartenir à chacune des classes avec la formule suivante, pour un modèle probabiliste θ :

$$\begin{aligned} \pi(1|x) &= \sup_{\theta \in \Theta} \min[\pi_\Theta(\theta), p_\theta(1|x) - p_\theta(0|x)], \\ \pi(0|x) &= \sup_{\theta \in \Theta} \min[\pi_\Theta(\theta), p_\theta(0|x) - p_\theta(1|x)], \end{aligned} \tag{6}$$



(a) Échantillon (b) Incertitude

(c) Aléatoire: $\mathcal{U}_a$ (d) Ép.: $\mathcal{U}_e$ (e) Totale: $\mathcal{U}$

Figure 2: Échantillon avec zones d'incertitudes classiques, épistémiques et aléatoires

avec $\pi_\Theta(\theta)$ dépendant de la vraisemblance $L(\theta)$ et du maximum de vraisemblance $L(\hat{\theta})$:

$$\pi_\Theta(\theta) = \frac{L(\theta)}{L(\hat{\theta})}. \tag{7}$$

L'incertitude épistémique est alors élevée lorsque les deux classes sont très plausibles, tandis que l'incertitude aléatoire est élevée lorsque les deux classes ne sont pas plausibles :

$$\begin{aligned} \mathcal{U}_e(x) &= \min[\pi(1|x), \pi(0|x)], \\ \mathcal{U}_a(x) &= 1 - \max[\pi(1|x), \pi(0|x)]. \end{aligned} \tag{8}$$

Ce calcul dépend non seulement de la prédiction du modèle, mais aussi des observations. En résumé, moins il y a d'observations dans une région, ou moins il y a d'éléments de décision pour prédire fortement une classe, plus la plausibilité des deux classes est élevée et plus l'incertitude est réductible (et donc épistémique) par l'ajout de connaissances.

Un exemple est présenté sur la figure 2, un jeu de données à deux classes est présenté sur 2a et les zones d'incertitude du modèle sont présentées sur 2b selon l'échantillonnage par incertitude (4)-(5). Une ligne horizontale peut être distinguée là où l'incertitude du modèle est la plus élevée. Cependant, l'échantillon représenté dans 2a, montre qu'une partie de l'incertitude peut être éliminée plus facilement en ajoutant des observations. Dans la même figure, trois jeux de données différents montrent comment l'échantillon peut évoluer en ajoutant des observations. Quelle que soit la distribution finale, l'incertitude à gauche n'est pas très réductible, tandis que l'incertitude à droite peut être modifiée par l'ajout de connaissances.

Ces deux incertitudes peuvent être calculées à l'aide de l'équation (8), et sont illustrées sur la même figure. L'incertitude aléatoire est représentée par 2c et l'incertitude épistémique par 2d. L'incertitude totale est alors la somme des deux 2e.

Ces informations peuvent être utiles pour trouver des zones d'incertitude réductible, mais elles ne sont pas compatibles avec des étiquettes plus riches qui contiennent également de l'incertitude. La méthode de calcul dépend également des observations en plus du modèle. En outre, le problème de l'exploration-exploitation n'est pas entièrement traité. Ceci conduit à la section suivante dans laquelle une stratégie d'échantillonnage par incertitude pour les étiquettes riches est proposée, elle est également étendue à plusieurs classes.

4 Étiquettes riches et classes multiples

Dans cette section, nous proposons une stratégie d'échantillonnage par incertitude, avec une phase de calcul simplifiée, capable de traiter des étiquettes plus riches et ne dépendant plus directement des observations, mais seulement de la prédiction du modèle. La méthode utilise la discorde et la non-spécificité et traite le problème de l'exploration-exploitation.

À partir d'ici, une étiquette peut être incertaine et imprécise. Pour la représentation utilisée dans ce document, plus le point est foncé, moins l'étiquette contient d'ignorance (*i.e. Je suis sûr qu'il s'agit d'un chien*), plus le point est clair, plus elle contient d'ignorance (*i.e. je n'ai aucune idée entre un chien et un chat*). Il est important de noter que les étiquettes ne sont plus "dures", mais modélisées par une fonction de croyance, qui permet une telle représentation.

4.1 Discorde et non-spécificité: Incertitude de Klir

La discorde et la non-spécificité sont des outils qui permettent de modéliser l'incertitude, nous utilisons la représentation de (Klir et Wierman, 1998) pour un échantillonnage par incertitude.

Discorde : Appliquée à la sortie d'un modèle capable de faire une prédiction incertaine et imprécise[2], elle représente les contradictions dans la prédiction et se calcule avec :

$$D(m) = -\sum_{A \subseteq \Omega} m(A) \log_2(BetP(A)), \tag{9}$$

avec m une fonction de masse, ou la sortie du modèle (voir section 2.2).

Non-spécificité : Elle permet de quantifier le degré d'ignorance du modèle, plus elle est élevée, plus la réponse du modèle est imprécise, elle se calcule avec :

$$N(m) = \sum_{A \subseteq \Omega} m(A) \log_2(|A|). \tag{10}$$

Incertitude de Klir : Elle est ensuite dérivée de la discorde et de la non-spécificité, en additionnant les deux formules précédentes :

$$\mathcal{U}_m(x) = N(x) + D(x), \tag{11}$$

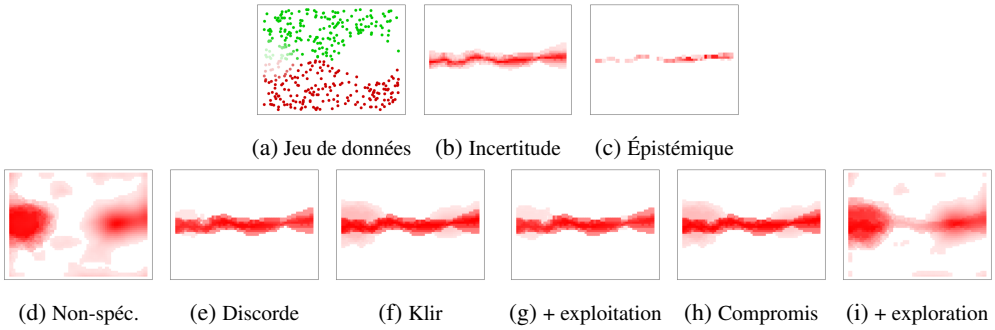
avec $N(x)$ et $D(x)$ respectivement la non-spécificité et la discorde du modèle en x. (Klir et Wierman, 1998) proposent d'utiliser le même poids pour la discorde et la non-spécificité, mais (Denoeux et Bjanger, 2000) introduisent un paramètre $\lambda \in [0, 1]$ permettant de donner plus de poids à la non-spécificité (plus d'exploration) ou à la discorde (plus d'exploitation) :

$$\mathcal{U}_m(x) = \lambda N(x) + (1 - \lambda) D(x). \tag{12}$$

Notons que cette incertitude est naturellement étendue à un nombre de classes $|\Omega| \geq 2$.

Cette formule présente l'avantage d'identifier l'incertitude totale ainsi que l'incertitude réductible, mais aussi de prendre en compte l'incertitude déjà présente dans les étiquettes et d'être ajustable pour plus d'exploration ou d'exploitation. La figure 3a montre un jeu de données avec deux zones d'incertitude, à droite une zone avec un manque de données et à gauche

[2]Le modèle *Evidential K-nearest Neighbors* de (Denoeux, 1995) est pris en compte.



(a) Jeu de données (b) Incertitude (c) Épistémique

(d) Non-spéc. (e) Discorde (f) Klir (g) + exploitation (h) Compromis (i) + exploration

Figure 3: Un jeu de données imparfaitement étiquetées (incertitude classique, incertitude épistémique et la proposition) et possibilité de faire un compromis d'exploration-exploitation.

une zone où les étiquettes sont plus ignorantes. L'échantillonnage par incertitude, (4) ou (5), n'est capable de voir aucune de ces deux zones 3b. L'incertitude épistémique (8) est capable de distinguer l'incertitude liée à la disposition des observations dans l'espace, mais pas l'incertitude liée à l'ignorance des sources 3c.

La proposition permet de représenter chacune de ces incertitudes. La deuxième ligne montre les zones de non-spécificité 3d, de discorde 3e et d'incertitude de Klir 3f.

Il est également possible de faire varier le résultat pour plus d'exploration ou plus d'exploitation en modifiant λ. Les dernières figures montrent les zones d'incertitude pour différentes valeurs de λ, plus de discorde à gauche 3g à plus de non-spécificité à droite 3i

5 Discussion et Conclusion

Le calcul de l'incertitude épistémique est exigeant, et pas nécessairement accessible. Il est, selon les observations, nécessaire de passer par plusieurs phases de calcul, d'estimation de la vraisemblance, de maximum de vraisemblance et d'optimisation. Dans cet article, nous avons proposé une nouvelle stratégie d'échantillonnage par incertitude. L'objectif est de prendre en compte l'incertitude présente dans les étiquettes plus riches, ce qui n'était pas possible jusqu'à présent. La stratégie s'inspire de l'incertitude de Klir, combinant la discorde et la non-spécificité dans la sortie du modèle.

La simplicité a évidemment une contrepartie : le modèle doit être capable de fournir une fonction de masse, pour représenter l'incertitude et l'imprécision dans le résultat. De tels modèles existent, mais ne sont pas nombreux, parmi lesquels on trouve les K plus proches voisins crédibilistes (Denoeux, 1995), les arbres de décision crédibilistes (Hoarau et al., 2023), les forêts aléatoires crédibilistes et même les réseaux de neurones crédibilistes.

La nouveauté de ce travail réside dans la représentation de nouvelles informations en échantillonnage par incertitude, plutôt que dans la comparaison des performances. L'étape suivante consiste à appliquer ces modèles à l'apprentissage actif, où le modèle d'apprentissage a accès à un nombre très limité d'observations étiquetées et doit choisir les observations les plus pertinentes à étiqueter afin d'augmenter les performances. La capacité du modèle à définir des zones d'incertitude et à catégoriser ces incertitudes constitue alors une information pertinente.

Références

Aggarwal, C., X. Kong, Q. Gu, J. Han, et P. Yu (2014). *Active Learning : A Survey, Data Classification : Algorithms and Applications*. CRC Press.

Bondu, A., V. Lemaire, et M. Boullé (2010). Exploration vs. exploitation in active learning : A bayesian approach. In *International Joint Conference on Neural Networks*.

Dempster, A. P. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics 38*(2), 325 – 339.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *Systems, Man and Cybernetics, IEEE Transactions on 219*.

Denoeux, T. et M. Bjanger (2000). Induction of decision trees from partially classified data using belief functions. *Systems, Man, and Cybernetics, IEEE International Conference*.

Hoarau, A., A. Martin, J.-C. Dubois, et Y. Le Gall (2023). Evidential random forests. *Expert Systems with Applications 230*.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety 54*(2), 217–223. Treatment of Aleatory and Epistemic Uncertainty.

Hüllermeier, E. et W. Waegeman (2021). Aleatoric and epistemic uncertainty in machine learning : An introduction to concepts and methods. *Machine Learning 110*, 457–506.

Kendall, A. et Y. Gal (2017). What uncertainties do we need in bayesian deep learning for computer vision ? In *NIPS*.

Klir, G. J. et M. J. Wierman (1998). Uncertainty-based information : Elements of generalized information theory. In *Springer-Verlag*.

Nguyen, V.-L., M. H. Shaker, et E. Hüllermeier (2022). How to measure uncertainty in uncertainty sampling for active learning. *Mach. Learn. 111*, 89–122.

Senge, R., S. Bösner, K. Dembczynski, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, et E. Hüllermeier (2014). Reliable classification : Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Inf. Sci. 255*, 16–29.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton : Princeton University Press.

Smets, P. et R. Kennes (1994). The transferable belief model. *Artificial Intelligence 66*(2).

Summary

Recent research in active learning, and more precisely in uncertainty sampling, has focused on the decomposition of model uncertainty into reducible and irreducible uncertainties. In this paper, we propose to simplify the computational phase and remove the dependence on observations, but more importantly to take into account the uncertainty already present in the labels, *i.e.* the uncertainty of the oracles. The proposed strategy, sampling by Klir uncertainty, addresses the exploration-exploitation dilemma using the theory of belief functions.