Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 243–246
DOI: 10.15439/2022F18
ISSN 2300-5963 ACSIS, Vol. 30

Three-way Learnability: A Learning Theoretic Perspective on Three-way Decision

Andrea Campagner, Davide Ciucci
Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano–Bicocca, Viale Sarca 336/14, 20126 Milano, Italy

***Abstract*—In this article we study the theoretical properties of Three-way Decision (TWD) based Machine Learning, from the perspective of Computational Learning Theory, as a first attempt to bridge the gap between Machine Learning theory and Uncertainty Representation theory. Drawing on the mathematical theory of orthopairs, we provide a generalization of the PAC learning framework to the TWD setting, and we use this framework to prove a generalization of the Fundamental Theorem of Statistical Learning. We then show, by means of our main result, a connection between TWD and selective prediction.**

I. Introduction

In the recent years, there has been an increasing interest toward exploring the connections between learning theory and different uncertainty representation theories: This trend includes both the generalization of standard learning-theoretic tools and techniques to settings that involve representation formalisms that are more general than probability theory [1], [2], as well as the theoretical study of algorithms inspired by uncertainty representation [3], [4].

Among other uncertainty representation theories, Three-way decision (TWD) is an emerging computational paradigm, first proposed by Yao in Rough Set Theory [5], based on the simple idea of *thinking in three "dimensions"* (rather than in binary terms) when representing and managing computational objects [6]: in the Machine Learning (ML) [7] setting, this notion is usually declined in terms of allowing ML models to *abstain*. This approach attracted a large interest, also justified by promising empirical results in different ML tasks such as active learning [8], [9], cost-sensitive classification [10], clustering [11], [12], [9]. Despite these promising empirical results, the theoretical foundations of TWD-based ML received so far little attention [13], [14]. Indeed, even though, in the recent years, there has been an increasing interest toward generalizing computation learning theory (CLT) to cautious inference methods such as *selective prediction* [15] or the KWIK (*Knows what it Knows*) framework [16], such results cannot be easily applied to the TWD setting: While in the TWD setting abstention is a property of single classifiers; in the latter two frameworks abstention is usually achieved by consensus voting.

In this article, we study the generalization of a standard CLT mathematical framework, the so-called *Probably Approximately Correct* (PAC) learning framework, to the TWD setting: In particular, we will provide a generalization of the *Fundamental Theorem of Learning* to the TWD setting, and we show that our result generalizes previous results in the selective prediction setting. More in detail, the rest of this article is structured as follows: In Section II we provide the necessary mathematical background on TWD (in Section II-A) and CLT (in Section II-B); in Section III we describe the generalization of the PAC learning framework to the TWD setting and we prove our main result; finally, in Section IV, we summarise our contribution and describe possible research directions.

II. Background

A. Three-way Decision and Orthopairs

In this work we will refer to the formalization of TWD-based ML models (in the following, TW Classifiers) as *orthopairs*:

Definition 1. *An* orthopair *[17] over the universe X (which represents the instance space) is a pair of sets $O = (P, N)$ such that $P, N \subseteq X$ and $P \cap N = \emptyset$, with P and N standing, respectively, for* positive *and* negative. *The* boundary *is defined as $Bnd = (P \cup N)^c$.*

An orthopair represents an uncertain concept: Specifically, the status of the elements in the boundary is uncertain (i.e., it is not known whether they belong to the concept). Thus, a given orthopair stands as an approximation for a collection of consistent concepts:

Definition 2. *We say that an orthopair $O = (P, N)$ is* consistent *with a concept $C \subset X$ if $x \in P \implies x \in C$ and $x \in N \implies x \notin C$.*

Finally, we remark that it is possible to define different orderings between orthopairs: In particular, O_2 is *less informative* than O_1, denoted $O_2 \leq_I O_1$ if $P_2 \subseteq P_1$ and $N_2 \subseteq N_1$.

B. Computational Learning Theory

Computational Learning Theory [18] (CLT) refers to the branch of Machine Learning and Theoretical Computer Science focusing on the theoretical study of learning algorithms. Various mathematical formalisms have been proposed toward this goal, in this article we will refer to the PAC (probably approximately correct) learning framework, first proposed in [19]. Formally, let X be the instance space and Y be the target space, in this article we will focus on the *binary classification* setting, that is $Y = \{0, 1\}$. We assume that the observable data is generated i.i.d. according to an unknown probability distribution $\mathcal{D}$ over $X \times Y$. Let $\mathcal{H}$ be a hypothesis

class, that is a collection of functions $h : X \mapsto Y$, we define the *true risk* of h w.r.t. $\mathcal{D}$ as:

$$\mathcal{L}_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}\left[l(h(x), y)\right] = \int_{X \times y} l(h(x), y) d\mathcal{D}(x, y) \quad (1)$$

where $l : Y^2 \mapsto \mathbb{R}^+$ is a loss function. Since $\mathcal{D}$ is unknown, the true risk cannot be computed: It is usually approximated through the so-called *empirical risk* based on a sample, called *training set*, $S = (\langle x_1, y_1\rangle, \ldots, \langle x_m, y_m\rangle)$:

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i) \quad (2)$$

Given a training set S, we denote by S_X the tuple $S_X = (x_1, ..., x_m)$, and by S_Y the tuple $S_Y = (y_1, ..., y_m)$. The *Empirical Risk Minimization* w.r.t. the hypothesis class $\mathcal{H}$ is the family of algorithms $ERM_{\mathcal{H},m} : (X \times Y)^m \mapsto \mathcal{H}$ s.t. $ERM_{\mathcal{H},m}(S) \in argmin_{h \in \mathcal{H}} L_S(h)$, where $S = (\langle x_1, y_1\rangle, \ldots, \langle x_m, y_m\rangle)$ is the training set.

The *Fundamental Theorem of Learning* [20] establishes a relation between the true risk and empirical risk for the ERM algorithm w.r.t. a hypothesis class $\mathcal{H}$ which depends only on the so-called VC dimension, a combinatorial dimension of the complexity of $\mathcal{H}$.

Theorem 1. *Let $\mathcal{H}$ be a hypothesis class with VC dimension d. For each $\epsilon, \delta \in (0, 1)$ and distribution $\mathcal{D}$, then if $ERM_{\mathcal{H}}$ is given a dataset S of size $m \geq n_0$, with*

$$n_0 = O(\frac{d + ln(\frac{1}{\delta})}{\epsilon^2}) \quad (3)$$

with probability greater than $1 - \delta$, it holds that $|L_{\mathcal{D}}(ERM_{\mathcal{H}}(S)) - L_S(ERM_{\mathcal{H}}(S))| \leq \epsilon$. If, further, the realizability[1] *assumption holds, then, if S is a dataset of size $m \geq n_1$, with*

$$n_1 = O(\frac{d + ln(\frac{1}{\delta})}{\epsilon}) \quad (4)$$

with probability greater than $1 - \delta$, it holds that $L_{\mathcal{D}}(ERM_{\mathcal{H}}(S)) \leq \epsilon$.

Few works have studied the generalization of CLT results to hypothesis that can be described as orthopairs (that is, classifiers that can abstain on selected instances), mainly under the framework of *selective prediction* [21]: In this setting, the goal is to design learning algorithms $\mathcal{A}_{\mathcal{H},m} : (X \times Y)^m \mapsto \mathcal{O}_{\mathcal{H}}$, where $\mathcal{O}_{\mathcal{H}} \subseteq TW(\mathcal{H})$ (see Eq. (15)), s.t. $L_{\mathcal{D}}(A(S)) = 0$ but $A(S)$ is allowed to abstain on certain instances. This abstention is usually achieved either by the combination of a standard hypothesis $h : X \to Y$ with a rejection function $r : X \to \{\bot, \top\}$, or, equivalently, by consensus voting based on a version space $V \subseteq \mathcal{H}$ [21]. As we show in the following sections (specifically, in Section III-A) the setting we consider is a proper generalization of selective prediction. More recently, the application of orthopairs in CLT has been studied in the setting of adversarial machine learning [22], as well as to characterize the generalization capacity of hypothesis classes under generative assumptions [23]. We note, however, that even though the above mentioned work and the framework we study in this article rely on the representation formalism of orthopairs, the aims of these three frameworks are essentially orthogonal, also in terms of the mathematical techniques adopted: Indeed, while the three-way learning framework we study relies on a generalization of the ERM paradigm, the frameworks studied in [23], [22] rely on a transductive learning approach.

[1]Here realizability means that $\exists h \in \mathcal{H}$ s.t. $L_D(h) = 0$.

III. Three-way Learning

In this Section, we provide a first study of a generalization of standard Computational Learning Theory to the setting of TW Classifiers. As hinted in Section II-A, we will represent a TW Classifier as an orthopair O; then, a hypothesis space of TW Classifier will be represented as a collection $\mathcal{O}$ of orthopairs over X. In the TWD literature, the risk of a TW Classifier is usually evaluated by means of a cost-sensitive generalization of the 0-1 loss: $l_{TW}(O(x), y) = \begin{cases} 1 & O(x) \bot y \\ \lambda_a & x \in Bnd_O \\ 0 & otherwise \end{cases}$, where $\lambda_a \in [0, 0.5)$ is the cost of abstention, and $O(x) \bot y$ is the error case, that is $(x \in P_O \wedge y = 0) \vee (x \in N_O \wedge y = 1)$. Compared to the standard definition of risk adopted in the TWD literature we assume that the cost of error is always 1. Based on the loss function l_{TW} we can define both the true risk $\mathcal{L}^{TW}_{\mathcal{D}}$ and the empirical risk L^{TW}_S. Evidently, the risk of O can be decomposed as the sum of two functions:

$$\begin{aligned} \mathcal{L}^{TW}_{\mathcal{D}}(O) &= \mathbb{E}_{\mathcal{D}}\left[l_{TW}(O(x), y)\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}_{O(x) \bot y}\right] + \lambda_a \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}_{x \in Bnd_O}\right] \\ &= Pr_{x \sim \mathcal{D}}(O(x) \bot y) \\ &\quad + \lambda_a \cdot Pr_{x \sim \mathcal{D}}(x \in Bnd_O) \end{aligned} \quad (5)$$

The same decomposition can be similarly applied for the empirical risk. Let $\mathcal{E}_{\mathcal{D}}(O) = Pr_{x \sim \mathcal{D}}(O(x) \bot y)$ and $\mathcal{A}_{\mathcal{D}}(O) = \lambda_a \cdot Pr_{x \sim \mathcal{D}}(x \in Bnd_O)$. We denote with $\mathcal{O}^{OPT} = \{O \in \mathcal{O} : \mathcal{E}_{\mathcal{D}}(O) = min_{O' \in \mathcal{O}} \mathcal{E}_{\mathcal{D}}(O')\}$. We say that $\mathcal{D}$ is *weakly realizable* w.r.t. $\mathcal{O}$ if $\forall O^* \in \mathcal{O}^{OPT}$ it holds that $\mathcal{E}_{\mathcal{D}}(O^*) = 0$. If, furthermore, $\exists O^* \in \mathcal{O}^{OPT}$ s.t. $\mathcal{A}_{\mathcal{D}}(O^*) = 0$, then we say that $\mathcal{D}$ is *strongly realizable*. Through this article, we will assume only weak realizability. Compared to the realizability assumption, weak realizability assumption is indeed much weaker. As an example if the vacuous TW classifier $O_\bot = (\emptyset, \emptyset) \in \mathcal{O}$, then every distribution $\mathcal{D}$ is trivially weakly realizable w.r.t $\mathcal{O}$, while it is clearly not strongly realizable.

Let $\epsilon \in (0, 1)$, $\alpha \in (0, \lambda_a)$, then $O \in \mathcal{O}$ makes an (ϵ, α)-failure if one of the following holds:

$$\mathcal{E}_{\mathcal{D}}(O) > \epsilon, \quad \mathcal{A}_{\mathcal{D}}(O) > min_{O \in \mathcal{O}^{OPT}} \mathcal{A}_{\mathcal{D}}(O) + \alpha \quad (6)$$

Thus, O (ϵ, α)-fails if either its error is greater than ϵ, or if its abstention rate is greater, by a margin of at least α, than the lowest abstention rate among those TW Classifiers that make no error. We thus define the notion of *Three-way learnability*:

Definition 3. *$\mathcal{O}$ is* Three-way learnable *if exists an algorithm $C_m : (X \times Y)^m \mapsto \mathcal{O}$ and $m_{\mathcal{O}} : (0, 1)^2 \times (0, \lambda_a) \mapsto \mathbb{N}$ such*

that, for each distribution $\mathcal{D}$, $\epsilon \in (0,1), \delta \in (0,1), \alpha \in (0, \lambda_a)$ $\forall m \geq m_{\mathcal{O}}(\epsilon, \delta, \alpha)$, and given $S \sim \mathcal{D}^m$, C returns $O \in \mathcal{O}$, s.t. O (ϵ, α)-fails with probability lower than δ

We then want to provide a characterization for TW learnability, similar to Theorem 1. For this purpose, we first define a generalization of the ERM algorithm to the TWD setting, that we call Three-way Risk Minimization (TW-RM):

Definition 4. *Let $S \in (X \times Y)^m$. Then,*

$$\begin{aligned} TWRM(S) = argmax_{O \in \mathcal{O}} \mathcal{A}_{X \setminus S_X}(O) \text{ s.t.} \\ \mathcal{E}_S(O) = min_{O' \in \mathcal{O}} \mathcal{E}_S(O') \\ \mathcal{A}_S(O) = min_{O' \in \mathcal{O}^{OPT}} \mathcal{A}_S(O') \end{aligned} \tag{7}$$

Thus, the TWRM algorithm selects, among those TW classifiers with minimal empirical risk, the TW classifier with maximal abstention rate on the non-observed instances (that is, the instances in $X \setminus S_X$). This has the goal of minimizing errors on non-observed instances, and is analogous to the *maximum margin* principle, and the *disagreement coefficient* in version space learning, active learning and selective prediction [15].

In order to characterize TW learnability, given hypothesis class $\mathcal{O}$ (i.e. a collection of orthopairs), we define two derived hypothesis classes. Given any orthopair $O \in \mathcal{O}$ we can define a classifier $h_O : X \mapsto \{0,1\}$, as: $h_O(x) = \begin{cases} 1 & x \in Bnd_O \\ 0 & otherwise \end{cases}$.

We denote the collection of such binary classifiers as $\mathcal{H}_{\mathcal{O}} = \{h_O : O \in \mathcal{O}\}$. Thus, given $\mathcal{O}$, the derived hypothesis class $\mathcal{H}_{\mathcal{O}}$ describes the abstention capacity of $\mathcal{O}$: In the classical setting $\mathcal{H}_{\mathcal{O}} = \{h_0\}$, where $\forall x \in X, \ \ h_0(x) = 0$, as no classifier in $\mathcal{O}$ is able to abstain: For all $O \in \mathcal{O}$, $Bnd_O = \emptyset$.

In regard to the second derived hypothesis class, we observe that the order $\leq_I$ defined in Section II-A defines a meet semi-lattice [17] on $\mathcal{O}$ with minimal element $O_\perp = (\emptyset, \emptyset)$. Then, we denote with $\mathcal{O}^\top = \{O \in \mathcal{O} : \nexists O' \in \mathcal{O} \text{ s.t } O \leq_I O'\}$, i.e. $\mathcal{O}^\top$ is the anti-chain of maximally informative elements of $\mathcal{O}$.

We now prove a generalization of Theorem 1 to the TWD setting, through which we show that the TW learnability of a hypothesis class $\mathcal{O}$, using the TWRM algorithm, can be characterized in terms of the derived hypothesis classes $\mathcal{H}_{\mathcal{O}}$ and $\mathcal{O}^\top$. In order to do so, we consider the VC dimension of the two derived hypothesis classes $\mathcal{H}_{\mathcal{O}}$ and $\mathcal{O}^\top$ as follows:

$$AVC(\mathcal{O}) = VC(\mathcal{H}_{\mathcal{O}}) \tag{8}$$

$$\begin{aligned} EVC(\mathcal{O}) = \sup\{|S| : S \subseteq X \wedge \forall C \subseteq S . \exists O \in \mathcal{O} \\ \text{s.t. } C = (P_O \cap S) \wedge (Bnd_O \cap S) = \emptyset\} \end{aligned} \tag{9}$$

Then, the following result holds:

Theorem 2. *Let $\mathcal{O}$ be s.t. $AVC(\mathcal{O}) = d_a$ and $EVC(\mathcal{O}) = d_e$. Then, for any distribution $\mathcal{D}$ weakly realizable w.r.t $\mathcal{O}$, $\epsilon, \delta \in (0,1), \alpha \in (0, \lambda_a)$, if $TWRM_{\mathcal{O}}$ is given a dataset of size m larger than :*

$$O\left(max\left\{\frac{1}{\epsilon}\left(d_e + ln\frac{1}{\delta}\right), \left(\frac{\lambda_a}{\alpha}\right)^2\left(d_a + ln\frac{1}{\delta}\right)\right\}\right) \tag{10}$$

then, $TWRM_{\mathcal{O}}(S)$ (ϵ, α)-fails with probability lower than δ.

Proof. We want to guarantee that the following bound holds:

$$\begin{aligned} Pr_{fail} = P(S : \exists O \in \mathcal{O} \ \ \wedge \\ |\mathcal{E}_D(O) - \mathcal{E}_S(O)| > \epsilon \ \ \vee \\ |\mathcal{A}_D(O) - \mathcal{A}_S(O)| > \alpha) < \delta \end{aligned} \tag{11}$$

Then, the results would follow by uniform convergence. By the union bound, it holds that:

$$\begin{aligned} Pr_{fail} \leq Pr(S : \exists O \in \mathcal{O}, |\mathcal{E}_D(O) - \mathcal{E}_S(O)| > \epsilon) \\ + Pr(S : \exists O \in \mathcal{O}, |\mathcal{A}_D(O) - \mathcal{A}_S(O)| > \alpha), \end{aligned} \tag{12}$$

thus, it is sufficient to jointly upper bound the two summands by $\frac{\delta}{2}$. As regards the error rate (i.e $\mathcal{E}$) bound, we note that:

$$\begin{aligned} Pr(S : \exists O \in \mathcal{O}, |\mathcal{E}_D(O) - \mathcal{E}_S(O)| > \epsilon) \\ Pr(S : \exists O \in \mathcal{O}^\top, \mathcal{E}_D(O) > \epsilon) \end{aligned} \tag{13}$$

Since $O^\top$ is a binary hypothesis class, then, by Theorem 1, the above bound holds with probability greater than $1 - \delta$ as long as $|S| \geq \frac{1}{\epsilon}\left(d_e + ln\frac{1}{\delta}\right)$. Furthermore, by uniform convergence this holds, in particular, for $TWRM_{\mathcal{O}}(S)$.

For the abstention part, the same line of reasoning can be applied, however, as we only assume weak realizability, only the result in Theorem 1 that applies to agnostic learning can be used. Then, as long as $|S| \geq \left(\frac{\lambda_a}{\alpha}\right)^2\left(d_a + ln\frac{1}{\delta}\right)$ it holds that $|\mathcal{A}_D(O) - \mathcal{A}_S(O)| < \alpha$ with probability greater than $1 - \delta$. This holds, in particular for $TWRM_{\mathcal{O}}(S)$, and thus the theorem follows by uniform convergence and Eq. (12). □

As a simple corollary, in the strong realizable setting, it can be easily verified that:

Corollary 1. *Let $\mathcal{O}$ be s.t. $AVC(\mathcal{O}) = d_a$ and $EVC(\mathcal{O}) = d_e$. Then, for any distribution $\mathcal{D}$ strongly realizable w.r.t $\mathcal{O}$, $\epsilon, \delta \in (0,1), \alpha \in (0, \lambda_a)$, if $TWRM_{\mathcal{O}}$ is given a dataset of size m larger than :*

$$O\left(max\left\{\frac{1}{\epsilon}\left(d_e + ln\frac{1}{\delta}\right), \frac{\lambda_a}{\alpha}\left(d_a + ln\frac{1}{\delta}\right)\right\}\right) \tag{14}$$

then, $TWRM_{\mathcal{O}}(S)$ (ϵ, α)-fails with probability lower than δ.

Note that, if $|\mathcal{O}| < \infty$, then it can be easily shown that $AVC(\mathcal{O}) \leq log_2(\mathcal{H}_{\mathcal{O}})$. Furthermore, it also holds that $EVC(\mathcal{O}) \leq log_2(\mathcal{O}^\top)$, as if O satisfies Eq. (8), then it obviously holds that $Bnd_O = \emptyset$ and hence $O \in \mathcal{O}^\top$.

A. Three-way Learning and Selective Prediction

Finally, we show that the proposed mathematical framework and the obtained results can be used to establish a connection between TWD and *selective prediction*. This result relies on the connection between version space theory and orthopairs [17], and allows us to derive a generalization bound, originally proven by El-Yaniv et al. [21], for selective prediction: This shows that the latter setting can be understood as a special case of TWD. Let $\mathcal{H}$ be a hypothesis class of binary classifiers, we call the Three-way Closure of $\mathcal{H}$, denoted as $TW(\mathcal{H})$, the hypothesis space obtained as:

$$TW(\mathcal{H}) = \bigcup\{O_H : H \subseteq \mathcal{H}\} \tag{15}$$

where, for each $H \subseteq \mathcal{H}$, $O_H = (\{x : \forall h \in H.h(x) = 1\}, \{x : \forall h \in H.h(x) = 0\})$. Basically, we associate with each possible version space H in $\mathcal{H}$ a corresponding orthopair O_H which abstains on every instance for which the hypotheses in H disagree [17]. Then we can prove the following result:

Corollary 2. *Let $|\mathcal{H}| < \infty$, let $\mathcal{O} = TW(\mathcal{H})$ the Three-way Closure of $\mathcal{H}$, and let $\lambda_a = 1$. Then, for any distribution $\mathcal{D}$ strongly realizable w.r.t $\mathcal{O}$, and for any $\delta \in (0,1)$, if $TWRM_{\mathcal{O}}$ is given a dataset of size m, then:*

1) *With probability 1 it holds that $\mathcal{E}_{\mathcal{D}}(TWRM_{\mathcal{O}}(S)) = 0$;*
2) *With probability greater than $1-\delta$ it holds that:*

$$\mathcal{A}_{\mathcal{D}}(TWRM_{\mathcal{O}}(S)) \leq O\left(\frac{1}{m} ln\left(\frac{|\mathcal{H}_{\mathcal{O}}|}{\delta}\right)\right) \quad (16)$$

$$= O\left(\frac{1}{m}\left(|\mathcal{H}| + ln\frac{1}{\delta}\right)\right) \quad (17)$$

Proof. The first equality easily follows from strong realizability and by noting that, by definition of $TW(\mathcal{H})$, $x \notin Bnd_{TWRM_{\mathcal{O}}(S)}$ iff $(x \in S_X \vee \exists v \in \{0,1\}.\forall h \in \{h' \in \mathcal{H} : \mathcal{E}_S(h) = 0\},\ h(x) = v)$. In regard to the second statement, the first inequality follows by standard algebraic manipulations. The equality, on the other hand, follows by noting that $|\mathcal{H}_{\mathcal{O}}| = 2^{\mathcal{H}}$ (as $TW(\mathcal{H})$ contains a TW classifier for each possible subset of hypotheses in $\mathcal{H}$). □

IV. Conclusion

In this article, we aimed at providing an initial study on the generalization of CLT results to the TWD setting. To this purpose, we first proposed an extension of the standard PAC learning framework to the TWD setting, that we called Three-way Learning and showed that our results generalize the previously known results in the selective prediction literature. As our results represent only a first direction in the theoretical study of TWD as applied to Machine Learning, we believe that the following questions would be of particular interest:

- Our analysis in Theorem 2 relies on a generalization of the VC dimension to the TWD setting. Tighter bounds can usually be obtained by relying on concepts such as Rademacher complexities or covering numbers [18]. How can these be generalized to TWD?
- In Corollary 2 we proved that, in the realizable case, selective prediction can be understood as a special case of TWD learning. Does this analysis also applies to the agnostic (i.e. non-realizable) setting [15]?
- PAC-Bayes bounds [24] study generalization bounds that apply when a probability distribution is defined over the hypothesis space. How can the PAC-Bayes framework be generalized to TWD? Interestingly, a very similar open problem has recently been posed also in Belief Function Theory (BFT) [25]. Due to the connection with random sets, a belief function can be seen as a probability distribution over orthopairs [26]: Then, the generalization of the PAC-Bayes framework to TWD would also enable studying the relationships between TWD and BFT.

References

[1] E. Hüllermeier, "Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization," *International Journal of Approximate Reasoning*, vol. 55, no. 7, pp. 1519–1534, 2014.

[2] G. Ma, F. Liu, G. Zhang, and J. Lu, "Learning from imprecise observations: An estimation error bound based on fuzzy random variables," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–8.

[3] S. Abbaszadeh and E. Hullermeier, "Machine learning with the sugeno integral: The case of binary classification," *IEEE Transactions on Fuzzy Systems*, 2020.

[4] E. Hüllermeier and A. F. Tehrani, "On the vc-dimension of the choquet integral," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2012, pp. 42–50.

[5] Y. Yao, "Three-way decision: an interpretation of rules in rough set theory," in *International Conference on Rough Sets and Knowledge Technology*. Springer, 2009, pp. 642–649.

[6] M. Ma, "Advances in three-way decisions and granular computing," *Knowl.-Based Syst*, vol. 91, pp. 1–3, 2016.

[7] M. Hu, "Three-way bayesian confirmation in classifications," *Cognitive Computation*, pp. 1–20, 2021.

[8] F. Min, S.-M. Zhang, D. Ciucci, and M. Wang, "Three-way active learning through clustering selection," *International Journal of Machine Learning and Cybernetics*, pp. 1–14, 2020.

[9] H. Yu, X. Wang, G. Wang, and X. Zeng, "An active three-way clustering method via low-rank matrices for multi-view data," *Information Sciences*, vol. 507, pp. 823–839, 2020.

[10] H. Li, L. Zhang, X. Zhou, and B. Huang, "Cost-sensitive sequential three-way decision modeling using a deep neural network," *International Journal of Approximate Reasoning*, vol. 85, pp. 68–78, 2017.

[11] M. K. Afridi, N. Azam, and J. Yao, "Variance based three-way clustering approaches for handling overlapping clustering," *International Journal of Approximate Reasoning*, vol. 118, pp. 47–63, 2020.

[12] P. Wang and Y. Yao, "Ce3: A three-way clustering method based on mathematical morphology," *Knowledge-based systems*, vol. 155, pp. 54–65, 2018.

[13] A. Campagner, F. Cabitza, P. Berjano, and D. Ciucci, "Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches," *Information Sciences*, vol. 579, pp. 347–367, 2021.

[14] A. Campagner and D. Ciucci, "A formal learning theory for three-way clustering," in *International Conference on Scalable Uncertainty Management*. Springer, 2020, pp. 128–140.

[15] R. Gelbhart and R. El-Yaniv, "The relationship between agnostic selective classification, active learning and the disagreement coefficient." *J. Mach. Learn. Res.*, vol. 20, no. 33, pp. 1–38, 2019.

[16] L. Li, M. L. Littman, T. J. Walsh, and A. L. Strehl, "Knows what it knows: a framework for self-aware learning," *Machine learning*, vol. 82, no. 3, pp. 399–443, 2011.

[17] D. Ciucci, "Orthopairs and granular computing," *Granular Computing*, vol. 1, no. 3, pp. 159–170, 2016.

[18] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[19] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[20] V. Vapnik, "On the uniform convergence of relative frequencies of events to their probabilities," in *Doklady Akademii Nauk USSR*, vol. 181, no. 4, 1968, pp. 781–787.

[21] R. El-Yaniv *et al.*, "On the foundations of noise-free selective classification." *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[22] S. Goldwasser, A. T. Kalai, Y. Kalai, and O. Montasser, "Beyond perturbations: Learning guarantees with arbitrary adversarial test examples," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 859–15 870, 2020.

[23] N. Alon, S. Hanneke, R. Holzman, and S. Moran, "A theory of pac learnability of partial concept classes," *arXiv preprint arXiv:2107.08444*, 2021.

[24] O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor, "Pac-bayes analysis beyond the usual bounds," *arXiv preprint arXiv:2006.13057*, 2020.

[25] F. Cuzzolin, *The geometry of uncertainty*. Springer, 2017.

[26] Y. Yao and P. Lingras, "Interpretations of belief functions in the theory of rough sets," *Information sciences*, vol. 104, no. 1-2, pp. 81–106, 1998.