Editing Boolean Classifiers: A Belief Change Perspective (extended version including the proofs of propositions)

Nicolas Schwind,[1] **Katsumi Inoue,** [2,3] **Pierre Marquis** [4,5]

[1] National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
[2] National Institute of Informatics, Tokyo, Japan
[3] The Graduate University for Advanced Studies, SOKENDAI, Tokyo, Japan
[4] Univ. Artois, CNRS, CRIL, F-62300 Lens, France
[5] Institut Universitaire de France
nicolas-schwind@aist.go.jp, inoue@nii.ac.jp, marquis@cril.fr

Abstract

This paper is about *editing* Boolean classifiers, i.e., determining how a Boolean classifier should be modified when new pieces of evidence must be incorporated. Our main goal is to delineate what are the rational ways of making such edits. This goes through a number of rationality postulates inspired from those considered so far for belief revision. We give a representation theorem and present some families of edit operators satisfying the postulates.

Introduction

Alice, a bank employee, receives Bob, a customer who wants to obtain a loan. Bob has a low income, but no debts. His record shows that he had already requested a loan in the past, and had fully reimbursed it. The bank management has recently provided Alice with an AI algorithm (a pre-trained predictor) to help her decide which issue to give to any loan application. Alice is asked to use this algorithm which recommends against granting Bob the requested loan due to the fact that he is not the owner of his principal residence. However, Alice is experienced and remembers of two customers Cindy and Dan with a profile similar to Bob's, who both had previously been granted a loan without any issue. Hence, Alice's expertise led her not to follow the recommendation of the AI algorithm and to grant Bob the loan requested. But Alice would like to do more to avoid that the problem encountered arises again with future clients having similar profiles. She wonders what could be done to this end.

The research question tackled in this paper is relevant to Alice's concern. We focus on Boolean classifiers φ: given an instance represented as a world, i.e., a truth assignment of all the variables of interest, φ classifies the instance as positive when it is a model of φ, and as negative when it is a counter-model of φ. The concept associated with φ is the set of all positive instances. Our very purpose is to determine how a Boolean classifier φ that has already been learned should be modified when new pieces of positive evidence / negative evidence μ (that may conflict with predictions of the classifier) are considered. We call such change operations on Boolean classifiers *positive edit / negative edits* (respectively), and we note them $\diamond^+$ and $\diamond^-$.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

We assume that both inputs φ and μ are represented as propositional formulae. Doing so, φ identifies both the classifier under consideration and, through its set of models, the associated concept. On the other hand, μ's models represent new positive (resp. negative) pieces of evidence in the case of a positive (resp. negative) edit. This representation choice allows one to deal with a number of existing ML classifiers: in an eXplainable AI perspective, many works have shown recently how ML classifiers C of various types can be associated with Boolean circuits φ_C, exhibiting the same input-output behaviours (i.e., the predictions made using C are precisely the same ones as those made using φ_C). The ML models that are concerned include not only decision trees (Izza, Ignatiev, and Marques-Silva 2020; Audemard et al. 2021) and decision lists (Ignatiev and Silva 2021), but also a number of ML models that are usually considered as less interpretable, like random forests (Audemard, Koriche, and Marquis 2020; Izza and Marques-Silva 2021), gradient boosted trees (Ignatiev 2020), some Bayes nets (Shih, Choi, and Darwiche 2018, 2019), and binary neural networks (Narodytska et al. 2018; Shi et al. 2020). Accordingly, classifiers C from those families can be taken into account in our framework, using φ_C as a representation of C since the two are prediction-equivalent.

Edit operations are connected to *incremental concept learners*, like Mitchell's Candidate Elimination Algorithm (Mitchell 1977), Schlimmer and Granger's `STAGGER` (Schlimmer and Fisher 1986), Fisher's `COBWEB` (Fisher 1987), and Gallant's Pocket Algorithm (Gallant 1988). Such systems, also referred to as *on-line learning systems*, are suited to learning scenarios when a whole training set is not available a priori but examples arrive over time. Borrowing the criteria used in (Maloof and Michalski 2000) to draw a typology of such systems, edit operations characterize on-line learning systems with (full) concept memory (the role played by the classifier φ), temporal batch (the set of models of μ is a new set of examples in the case of a positive edit, and a new set of counter-examples in the case of a negative edit), and no instance memory (the examples and counter-examples used to induce φ are not stored). However, previous work about incremental concept learners was typically centered on aspects that are not considered in this paper. These included the design of a number of on-line learners (based on specific concept representations, e.g., decision trees or decision rules), the evaluation of their empirical accuracy but also of their run-time efficiency (this can be a critical aspect since items in a data stream can be received at a so high rate that real-time guarantees are required to handle all of them (Domingos and Hulten 2000)), and finally the choice of examples that must be kept at each learning step (Maloof and Michalski 2004).

Contrastingly, in our work, the focus is on *an axiomatic approach*. We do not consider any specific concept representation, and do not make any assumption about how the batch of new examples or counter-examples are represented. We nevertheless suppose that the new piece of evidence μ that triggers the edit operation of φ is certain, i.e., not pervaded by any noise. Thus, stepping back to the loan scenario, Alice is sure that Bob should be granted the loan. Here, our main goal is to delineate the rational ways of making such edits. This goes through a number of rationality postulates.

To determine such postulates, we look back at the *core principles of belief revision* which aims to incorporate, in a rational way, a new piece of information into the belief set of an agent (Alchourrón, Gärdenfors, and Makinson 1985; Alchourrón and Makinson 1985; Gärdenfors 1988). The AGM postulates (for Alchourrón, Gärdenfors and Makinson 1985) aim to formalize a set of rationality conditions based on three main principles: *primacy of update* (the new information must be believed after the change), *consistency* (the resulting belief set must be kept consistent when the new information is consistent), and *minimal change* (if simply adding the new information to the belief set raises no conflict, then nothing else should be added or removed).

Adapting the postulates of belief revision into edit is not a trivial task: when the beliefs of an agent φ and the new information μ are represented by propositional formulae, when the conjunction of φ and μ is consistent, the revision of φ by μ corresponds to that conjunction (Katsuno and Mendelzon 1991). However, by representing a Boolean classifier φ and a set of incoming examples μ by two propositional formulae, one cannot reasonably require the edited classifier to be the conjunction of φ and μ whenever consistent: this process would unconditionally remove positive instances not explicitly questioned by μ, while also not incorporating the examples from μ previously classified negatively by φ.

Edit differs from belief revision in that the objects under consideration (all of which being represented by propositional formulae) are nevertheless of different nature. Thus, an agent's beliefs (represented by φ) correspond to a set of possible worlds to whom the one actual "true" world is believed to belong, while in edit, it makes perfect sense for several instances to be both members of the concept represented by a Boolean classifier. Likewise, every Boolean classifier φ is essentially "consistent": when φ has no model, it simply represents the empty concept. This explains also why the consistency principle is irrelevant to an edit operation.

Nevertheless, the primacy of update and minimality of change principles can be adapted to the edit context. For this purpose, after some formal preliminaries, we introduce in the edit postulates in the context of *positive* edit first (incorporating a batch of positive instances into a Boolean classifier). We also give a representation theorem and present some examples of positive edit operators. Then, we show how these postulates can be adapted to the case of a *negative* edit (i.e., where the arriving batch is interpreted as a set of negative instances) and make precise how a correspondence between the two operations can be formalized through a duality result. We then consider the case of a *full edit*, where both positive and negative instances can be considered in the same batch. Lastly, related work is discussed just before the conclusion. Proofs are given in an appendix.

Formal Preliminaries

We consider a propositional language $\mathcal{L}_{PS}$ built from a finite set PS of variables and the standard connectives. A *world* is a truth assignment of all variables from PS. The set of all worlds is denoted by Ω, and the set of models of a propositional formula $\varphi \in \mathcal{L}_{PS}$ (i.e., the set of worlds that make φ true) is denoted by $[\varphi]$. Given two formulae α, β, we write $\alpha \models \beta$ whenever $[\alpha] \subseteq [\beta]$ and $\alpha \equiv \beta$ when $[\alpha] = [\beta]$.

Belief revision aims to incorporate into the beliefs of an agent (a formula φ) a new piece of information (a formula μ). Thus a revision operator $\circ$ associates formulae φ, μ with a revised formula $\varphi \circ \mu$, and is expected to satisfy a set of rationality postulates:[1]

Definition 1 (KM revision operator)**.** *A revision operator $\circ$ is said to be a* KM revision operator *if it satisfies the following postulates:*

(R1) $\varphi \circ \mu \models \mu$
(R2) *If* $[\varphi \wedge \mu] \neq \emptyset$*, then* $\varphi \circ \mu \equiv \varphi \wedge \mu$
(R3) *If* $[\mu] \neq \emptyset$*, then* $[\varphi \circ \mu] \neq \emptyset$
(R4) *If* $\varphi \equiv \varphi'$ *and* $\mu \equiv \mu'$*, then* $\varphi \circ \mu \equiv \varphi' \circ \mu'$
(R5) $(\varphi \circ \mu) \wedge \mu' \models \varphi \circ (\mu \wedge \mu')$
(R6) *If* $[(\varphi \circ \mu) \wedge \mu'] \neq \emptyset$*, then* $\varphi \circ (\mu \wedge \mu') \models (\varphi \circ \mu) \wedge \mu'$

(R1) is the success postulate, it relates to the primacy of update principle: the new information must be believed after revision. **(R3)** is the consistency postulate. **(R4)** is the syntax-irrelevance postulate. And **(R2)**, **(R5)** and **(R6)** express the minimality of change conditions. We refer the reader to (Alchourrón, Gärdenfors, and Makinson 1985; Katsuno and Mendelzon 1991) for a deeper discussion about the rationale of these postulates.

Positive Edit

We now intend to define a change operation $\diamond^+$ that consists in editing an (already learned) Boolean classifier φ according to a new information μ. We assume that φ is represented by a propositional formula. In this context, each world represents an *instance*, and a world ω is a model of φ if and only if it is a positive instance of the concept represented by φ (so each instance is either classified as positive or negative by φ). The new information μ is called a *positive dataset* and is also represented by a propositional formula. The set of models of μ represents a batch of arriving positive instances, also

[1]We give here the KM postulates (Katsuno and Mendelzon 1991), which are the translation of the AGM postulates in finite propositional logic.

called positive *examples* (i.e., when referring to the models of μ). We do not make any further assumption on the way φ and μ are represented (e.g., φ could be a decision tree and μ a DNF formula, but it does not have to be the case).

Example 1. *Let us formalize the scenario provided in the introduction. We set $PS = \{p, q, r, s\}$ where p means that the applicant "has a high income", q stands for "owns her principal residence", r means "has no debts", and s means "has reimbursed a previous loan". We assume that $\varphi = p \wedge q \wedge r$, i.e., the predictor recommends granting a loan precisely to those residence owners having a high income and no debts. Then, let Bob have the profile $\omega_1 = 0011$, i.e., he is not owning his residence, has a low income, but has no debts and has reimbursed a previous loan; and let Cindy and Dan be identified with the same profile $\omega_2 = 0101$. The positive dataset μ is then defined as any propositional formula such that $[\mu] = \{\omega_1, \omega_2\}$, e.g., $\mu = \neg p \wedge s \wedge (q \leftrightarrow \neg r)$.*

An edit operator $\diamond^+$ associates every Boolean classifier φ and every positive dataset μ with an edited Boolean classifier $\varphi \diamond^+ \mu$. Our key assumption is that the new piece of evidence μ that triggers the edit operation of φ is provided by a domain expert: it is therefore certain, i.e., not pervaded by any noise. This can be ensured in a number of scenarios (thus, stepping back to the example given in the introduction, Alice is sure that applicants with the same profiles as Bob, Cindy and Dan should be granted the loan).

We are ready to introduce our postulates for positive edit:

Definition 2 (Positive Edit operator)**.** *An operator $\diamond^+$ is said to be a* positive edit operator *(PE operator for short) if it satisfies the following postulates:*

- ***(P1)*** $\mu \models \varphi \diamond^+ \mu$
- ***(P2)*** *If $\mu \models \varphi$, then $\varphi \diamond^+ \mu \equiv \varphi$*
- ***(P3)*** *If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$, then $\varphi_1 \diamond^+ \mu_1 \equiv \varphi_2 \diamond^+ \mu_2$*
- ***(P4)*** *If $\psi \models \varphi \diamond^+ \mu$, then $\varphi \diamond^+ \mu \equiv \varphi \diamond^+ (\mu \vee \psi)$*

(P1) relates to the primacy of update principle. Since the incoming positive dataset μ is assumed to be certain, **(P1)** requires the edited classifier to "comply" with μ, i.e., to correctly classify all examples from μ as positive instances. This can be viewed as the counterpart of **(R1)** in belief revision. **(P2)** is a minimality of change postulate: if the initial classifier already complies with μ, then there is no need to change it. It is reminiscent to **(R2)** in belief revision, but **(P2)** and **(R2)** differ in their premise. Indeed, **(P2)** does not say anything when $\mu \not\models \varphi$ and $[\varphi \wedge \mu] \neq \emptyset$: when φ does not comply with μ (i.e., some positive examples from μ were previously classified as negative instances by φ), then it makes perfect sense to question the concept membership of any instance $\omega \notin [\mu]$. Note that when $\mu \models \varphi$, the conclusion of **(P2)** can equivalently be stated as $\varphi \diamond^+ \mu \equiv \varphi \vee \mu$, from which the similarity with **(R2)** is clearer: μ is simply "added" to φ, which results in not changing φ at all. **(P3)** is the syntax-independence postulate, which is the direct counterpart of **(R4)**. **(P4)** is another minimality of change postulate. Its counterparts in belief revision are **(R5)** and **(R6)**, which together express that if φ revised by a first piece of information μ_1 is consistent with another piece of information μ_2, then revising φ by both pieces of information taken together (i.e., by $\mu_1 \wedge \mu_2$) boils down to "adding" μ_2 to the revision of φ by μ_1. Likewise, in our setting, **(P4)** says that if the edit of a classifier φ by a first positive dataset μ complies with another positive dataset ψ, then its edit by the two batches taken together (i.e., by $\mu \vee \psi$) boils down to "adding" ψ to the edit of φ by μ: indeed, if $\psi \models \varphi \diamond^+ \mu$, then $\varphi \diamond^+ \mu \equiv (\varphi \diamond^+ \mu) \vee \psi$ and thus the conclusion of **(P4)** can equivalently be written as $(\varphi \diamond^+ \mu) \vee \psi \equiv \varphi \diamond^+ (\mu \vee \psi)$. Please note that **(R3)**, the consistency postulate in belief revision, is the only postulate with no counterpart in our setting, since every Boolean classifier is essentially "consistent": if $[\varphi] = \emptyset$, then φ characterizes an empty concept.

Notably, the edit postulates **(P1-P4)** are also reminiscent to properties that can be sought for incremental concept learners in the absence of noise. Thus, **(P1)** states that once the edit operation has been performed, the resulting concept $\varphi \diamond^+ \mu$ must be consistent (in the sense of (Mitchell 1982)) with the new examples given by μ, which precisely means that those examples must be positive instances of $\varphi \diamond^+ \mu$. **(P2)** requires not to change the concept φ when it classifies correctly the new examples given by μ. This condition is achieved, for instance, by the perceptron update rule (Rosenblatt 1958). **(P3)** requires syntax not to play any role in the on-line learning process, which makes sense if the specific representation φ of the concept at hand is irrelevant (this is one of our starting assumptions). Finally, provided that **(P2)** holds, **(P4)** can be viewed as a relaxation of an order-independence condition that is satisfied by some on-line learners. This last property roughly states that while the new examples arrives over time, once the whole input sequence has been processed, the classifier has been transformed in the same way as if all pieces of evidence were available as a whole. Such an order-independence condition is ensured by `ID5` (Utgoff 1989), that has been shown to compute the same decision tree as the one that would be generated by `ID3` (Quinlan 1986), provided that the whole set of examples would be available at start. Formally, in our setting, the order-independence condition can be stated as $(\varphi \diamond^+ \mu) \diamond^+ \psi \equiv \varphi \diamond^+ (\mu \vee \psi)$. Note that this condition is quite demanding and not satisfied by every on-line learner (e.g., the perceptron update rule may easily question the way an instance μ has been classified when editing further the linear classifier by taking a new instance ψ into account). Accordingly, we focused on a weaker condition (it is easy to show that **(P4)** is a logical consequence of the order-independence condition when **(P2)** holds).

At this point, one can already identify a few simple operators from the class:

Definition 3 (Some PE operators)**.** *The* trivial*,* basic*, and* drastic *operators, respectively noted $\diamond^+_T$, $\diamond^+_B$, and $\diamond^+_D$, are defined for each classifier φ and each positive dataset μ as:*

- $\varphi \diamond^+_T \mu = \varphi$ *if* $\mu \models \varphi$, *otherwise* $\varphi \diamond^+_T \mu = \top$
- $\varphi \diamond^+_B \mu = \varphi \vee \mu$
- $\varphi \diamond^+_D \mu = \varphi$ *if* $\mu \models \varphi$, *otherwise* $\varphi \diamond^+_D \mu = \mu$

The trivial operator $\diamond^+_T$ just requires a classifier to classify all worlds as positive instances as soon as it does not initially comply with the input positive dataset. The basic operator $\diamond_B^+$ is simply defined as the disjunction: it simply adds as positive instances all the examples provided by μ which were not already classified as positive. The drastic operator $\diamond_D^+$ leaves the classifier φ unchanged if already compliant with μ. Otherwise, similarly to the trivial operator, it "forgets" everything, but classifies as positive instances precisely the ones explicitly provided as examples by μ. This can be viewed as the PE counterpart of the drastic revision operator $\circ_D$ defined as $\varphi \circ_D \mu = \varphi \wedge \mu$ if $[\varphi \wedge \mu] \neq \emptyset$, otherwise $\varphi \circ_D \mu = \mu$.

It is quite easy to check that these operators satisfy **(P1-P4)** (the proof is direct):

Proposition 1. *$\diamond_T^+$, $\diamond_B^+$ and $\diamond_D^+$ are PE operators.*

A Representation Theorem. Let us now show how PE operators can be characterized in terms of so-called *positive assignments*:

Definition 4 (Positive assignment)**.** *A* positive assignment *is a mapping associating every classifier φ with a mapping $f_\varphi : \mathcal{P}(\Omega) \mapsto \mathcal{P}(\Omega)$, such that for all classifiers φ, φ' and all subsets of worlds $W, W' \in \mathcal{P}(\Omega)$, the following properties are satisfied:*

1. $W \subseteq f_\varphi(W)$
2. *If $W \subseteq [\varphi]$, then $f_\varphi(W) = [\varphi]$*
3. *If $\varphi \equiv \varphi'$, then $f_\varphi = f_{\varphi'}$*
4. *If $W \subseteq W'$ and $W' \subseteq f_\varphi(W)$, then $f_\varphi(W) = f_\varphi(W')$*

Proposition 2. *An operator $\diamond^+$ is a PE operator if and only if there is a positive assignment $\varphi \mapsto f_\varphi$ such that for each classifier φ and each positive dataset μ, $[\varphi \diamond \mu] = f_\varphi([\mu])$.*

This is a "strong" representation result, in the sense that different positive assignments define different PE operators.

Interestingly, every mapping f_φ satisfies the property of *idempotence*:

Proposition 3. *For each positive assignment $\varphi \mapsto f_\varphi$ and each $W \subseteq \Omega$, we have that $f_\varphi(f_\varphi(W)) = f_\varphi(W)$.*

Accordingly, a consequence of the PE postulates is that $(\varphi \diamond^+ \mu) \diamond^+ \mu \equiv \varphi \diamond^+ \mu$. This idempotence property reflects a very simple form of minimal change and is standard in belief change: it is satisfied by belief revision operators but also by other forms of change operations, e.g., contraction (Caridroit, Konieczny, and Marquis 2017).

Noteworthy, condition 4 corresponds to the condition of Irrelevance of Rejected Contracts (IRC) in matching theory (Hatfield and Milgrom 2005). In that context, this property requires the removal of rejected contracts not to affect a choice set, and is a necessary condition to guarantee the existence of stable allocations (Aygün and Sönmez 2013), without implying rationalizability[2](Yang 2020).

Distance-based PE operators. We now introduce two classes of operators, called *dilation* operators and *min-generalization* operators. These operators are parameterized by a distance between worlds, i.e., a mapping $d : \Omega \times \Omega \mapsto \mathbb{N}$ such that $d(\omega, \omega') = 0$ if and only if $\omega = \omega'$, and that satisfies the triangular inequality property, i.e., $d(\omega, \omega'') \leq d(\omega, \omega') + d(\omega', \omega'')$, for all worlds $\omega, \omega', \omega''$.

Let us start with dilation operators, whose definition is inspired from the notion of formula dilation from (Bloch and Lang 2002; Dalal 1988). Given a classifier φ such that $[\varphi] \neq \emptyset$, and an integer k, the *k-dilation* of φ w.r.t. d, denoted by $D_\varphi^d(k)$, is defined by $D_\varphi^d(k) = \{\omega \in \Omega \mid d(\omega, \varphi) \leq k\}$, where $d(\omega, \varphi) = \min\{d(\omega, \omega') \mid \omega' \models \varphi\}$.

Definition 5 (Dilation operator)**.** *The dilation operator $\diamond_{\mathsf{dil},d}^+$ induced by d is defined for each classifier φ and each positive dataset μ by $[\varphi \diamond_{\mathsf{dil},d}^+ \mu] = [\mu]$ if $[\varphi] = \emptyset$, otherwise $[\varphi \diamond_{\mathsf{dil},d}^+ \mu] = \arg\min_k(\{D_\varphi^d(k) \mid [\mu] \subseteq D_\varphi^d(k)\})$.*

A number of dilation operators can be defined depending on the choice of d. For instance, consider the Hamming distance between worlds, denoted by d_H, defined for all worlds $\omega, \omega' \in \Omega$ as $d_H(\omega, \omega') = \{x \in PS \mid \omega(x) \neq \omega'(x)\}$ (Dalal 1988). Then the Hamming-based dilation operator $\diamond_{\mathsf{dil},d_H}^+$ consists in k-dilating φ w.r.t. d_H where k is the least integer for which the resulting set of models includes every model of μ (see Example 1 below).

Let us now introduce the class of min-generalization operators. Given a distance d and a world ω, let $\leq_\omega^d$ is the total preorder over worlds induced by ω and d and defined by $\omega' \leq_\omega^d \omega''$ iff $d(\omega', \omega) \leq_\omega^d d(\omega'', \omega)$. Given a classifier φ such that $[\varphi] \neq \emptyset$, the set $\min([\varphi], \leq_\omega^d)$ denotes the set of models of φ that have a distance to ω which is minimal among all models of φ, i.e., $\min([\varphi], \leq_\omega^d) = \{\omega' \in [\varphi] \mid \forall \omega'' \in [\varphi], d(\omega', \omega) \leq d(\omega'', \omega)\}$.

Definition 6 (Min-generalization operator)**.** *The min-generalization operator $\diamond_{\mathsf{gen},d}^+$ induced by d is defined for each classifier φ and each positive dataset μ by $[\varphi \diamond_{\mathsf{gen},d}^+ \mu] = [\mu]$ if $[\varphi] = \emptyset$, otherwise $[\varphi \diamond_{\mathsf{gen},d}^+ \mu] = \{\omega \in \Omega \mid \exists \omega', \omega'' \in \Omega, \omega' \models \mu, \omega'' \in \min([\varphi], \leq_{\omega'}^d), d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')\}$.*

The min-generalization operator consists in considering as positive instances by $\varphi \diamond_{\mathsf{gen},d}^+ \mu$ every world ω that is "in-between" (w.r.t. d) a model ω' of μ and a model ω'' of φ that is among the closest ones (w.r.t. d) to ω'. When $d = d_H$, the min-generalization operator can be characterized using the *most specific generalization* (msg) of the worlds involved.[3] Let $\mathsf{msg}(\omega, \omega')$ be the term $\bigwedge_{x \in PS \mid \omega(x) = \omega'(x) = 1} x \wedge \bigwedge_{x \in PS \mid \omega(x) = \omega'(x) = 0} \neg x$. Then one can check that $\varphi \diamond_{\mathsf{gen},d_H}^+ \mu \equiv \bigvee\{\mathsf{msg}(\omega, \omega') \mid \omega \in [\mu], \omega' \in \min([\varphi], \leq_\omega^d)\}$.

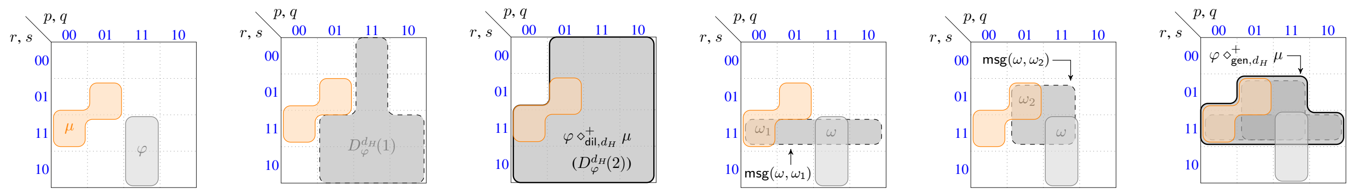
All the operators from these classes satisfy **(P1-P4)**:

Proposition 4. *For every distance d, the operators $\diamond_{\mathsf{dil},d}^+$ and $\diamond_{\mathsf{gen},d}^+$ are PE operators.*

Example 1 (continued)**.** *Let us go back to our loan scenario, and recall that $PS = \{p, q, r, s\}$, $\varphi = p \wedge q \wedge r$, and $\mu = \neg p \wedge s \wedge (q \leftrightarrow \neg r)$. Figure 1 depicts through*

[2]A choice function $\sigma : \mathcal{P}(E) \mapsto \mathcal{P}(E)$, i.e., a mapping such that $\sigma(S) \subseteq S$, is *rationalizable* if it can be characterized in terms of preference relation over elements from E.

[3]Most specific generalization is a key concept in machine learning, see e.g., (Plotkin 1970; Mitchell 1977).



(a) φ and μ (b) $D^{d_H}_{\varphi}(1)$ (c) $\varphi \diamond^{+}_{\mathsf{dil},d_H} \mu$ (d) $\mathsf{msg}(\omega, \omega_1)$ (e) $\mathsf{msg}(\omega, \omega_2)$ (f) $\varphi \diamond^{+}_{\mathsf{gen},d_H} \mu$

Figure 1: An example of Hamming-based dilation (Fig. 1c) and min-generalization (Fig. 1f) positive edits.

Karnaugh maps the models of φ and μ (Fig. 1a), the 1-dilation of φ (Fig. 1b), the Hamming-based dilation edit of φ by μ, which corresponds to the 2-dilation of φ (Fig. 1c), and the Hamming-based min-generalization edit of φ by μ (Fig. 1f), which corresponds to the disjunction of the two msgs given in Fig. 1d and 1e. Accordingly, we get that $\varphi \diamond^{+}_{\mathsf{dil},d_H} \mu \equiv p \vee q \vee r$ and $\varphi \diamond^{+}_{\mathsf{gen},d_H} \mu \equiv s \wedge (q \vee r)$.

As it can be verified on the example, dilation and min-generalization PE operators can easily add to φ models that are neither models of φ nor models of μ, thus questioning negative instances (the counter-models of φ). Such a generalization power (required by incremental learning) is not forbidden but not mandatory for PE operators (e.g., consider the basic and the drastic PE operators in Definition 3). Min-generalization PE operators may also question positive instances (again, this is expected when used for learning).

Negative Edit

Let us now consider the incorporation of negative instances into a Boolean classifier, alias negative edits. This time, the models of the change formula μ represent counter-examples of the target concept (a *negative dataset*).

Definition 7 (Negative Edit operator)**.** *An operator $\diamond^-$ is said to be a* negative edit operator *(NE operator for short) if it satisfies the following postulates:*

- ***(N1)*** $\mu \models \neg(\varphi \diamond^- \mu)$
- ***(N2)*** *If $\mu \models \neg\varphi$, then $\varphi \diamond^- \mu \equiv \varphi$*
- ***(N3)*** *If $\varphi_1 \equiv \varphi_2$ and $\mu_1 \equiv \mu_2$, then $\varphi_1 \diamond^- \mu_1 \equiv \varphi_2 \diamond^- \mu_2$*
- ***(N4)*** *If $\psi \models \neg(\varphi \diamond^- \mu)$, then $\varphi \diamond^- \mu \equiv \varphi \diamond^- (\mu \vee \psi)$*

Similarly to Harper and Levi's identities which show how a revision operator can be defined from a contraction operator and vice-versa (see e.g., (Caridroit, Konieczny, and Marquis 2017)), one can identify a correspondence between PE and NE operators. With an operator $\diamond^* : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$, let us associate an operator $\sigma(\diamond^*) : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$ defined as:

$$\varphi \, \sigma(\diamond^*) \, \mu = \neg(\neg\varphi \diamond^* \mu),$$

for every classifier φ and every formula μ. Then:

Proposition 5. *σ is an involution, that is, for each operator $\diamond^* : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$, $\sigma(\sigma(\diamond^*)) = \diamond^*$. Moreover, $\sigma(\diamond^*)$ is an NE operator if and only if $\diamond^*$ is a PE operator.*

We say that the operator $\sigma(\diamond^*)$ is the *dual* of $\diamond^*$. For instance, consider again the trivial, basic and drastic PE operators introduced in Definition 3. Then it is easy to see that the dual of these operators, i.e., the trivial, basic, and drastic NE operators, respectively denoted by $\diamond^-_T = \sigma(\diamond^+_T)$, $\diamond^-_B = \sigma(\diamond^+_B)$, and $\diamond^-_D = \sigma(\diamond^+_D)$, are defined for each classifier φ and each negative dataset μ by:

- $\varphi \diamond^-_T \mu = \varphi$ if $\mu \models \neg\varphi$, otherwise $\varphi \diamond^-_T \mu = \bot$
- $\varphi \diamond^-_B \mu = \varphi \wedge \neg\mu$
- $\varphi \diamond^-_D \mu = \varphi$ if $\mu \models \neg\varphi$, otherwise $\varphi \diamond^-_D \mu = \neg\mu$

Dual operators of dilation operators and min-generalization operators can also be easily defined. Interestingly, the operators dual to dilation operators involve an operation of formula *erosion* (Bloch and Lang 2002), which is an operation on formulae dual to the one of dilation. For instance, the Hamming-based erosion operator, denoted by $\diamond^-_{\mathsf{ero},d_H}$, is defined for each classifier φ and each negative dataset μ as $\varphi \diamond^-_{\mathsf{ero},d_H} \mu = \neg(\neg\varphi \diamond^+_{\mathsf{dil},d_H} \mu)$.

Full Edit

Let us finally consider the more general case when the new piece of evidence consists of both positive and negative instances that must be incorporated into the classifier. We call such a piece of evidence a *dataset*, i.e., a pair (μ^+, μ^-) such that μ^+ is a positive dataset (a set of examples), μ^- is a negative dataset (a set of counter-examples), and such that $[\mu^+ \wedge \mu^-] = \emptyset$. The set $\mathcal{D}$ denotes the set of all datasets.

Definition 8 (Full Edit operator)**.** *An operator $\diamond : \mathcal{L} \times \mathcal{D} \mapsto \mathcal{L}$ is said to be a* full edit operator *(FE operator for short) if for each a classifier φ and each dataset (μ^+, μ^-), it satisfies the following postulates:*

- ***(F1)*** $\mu^+ \models \varphi \diamond (\mu^+, \mu^-)$
- ***(F2)*** $\mu^- \models \neg(\varphi \diamond (\mu^+, \mu^-))$
- ***(F3)*** *If $\mu^+ \models \varphi$ and $\mu^- \models \neg\varphi$, then $\varphi \diamond (\mu^+, \mu^-) \equiv \varphi$*
- ***(F4)*** *If $\varphi_1 \equiv \varphi_2$, $\mu^+_1 \equiv \mu^+_2$ and $\mu^-_1 \equiv \mu^-_2$, then $\varphi_1 \diamond (\mu^+_1, \mu^-_1) \equiv \varphi_2 \diamond (\mu^+_2, \mu^-_2)$*
- ***(F5)*** *If $\psi \models \varphi \diamond (\mu^+, \mu^-)$ and $\alpha \models \neg(\varphi \diamond (\mu^+, \mu^-))$, then $\varphi \diamond (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$*

The postulate **(F1)** (resp. **(F2)**) corresponds to **(P1)** (resp. **(N1)**), while **(F3)** (resp. **(F4)**, **(F5)**) is a (weak) combination of **(P2)** and **(N2)** (resp. **(P3)** and **(N3)**, **(P4)** and **(N4)**).

A number of FE operators can be defined by means of a PE operator or an NE operator. Given an operator $\diamond^+ : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$, let us define the operator $\diamond_{\diamond^+} : \mathcal{L} \times \mathcal{D} \mapsto \mathcal{L}$ for each classifier φ and each dataset (μ^+, μ^-) as:

$$\varphi \diamond_{\diamond^+} (\mu^+, \mu^-) = (\varphi \diamond^+ \mu^+) \wedge \neg\mu^-.$$

We say that $\diamond_{\diamond^+}$ is *positively induced* by $\diamond^+$. Then:

Proposition 6. *$\diamond_{\diamond^+}$ is an FE operator if and only if $\diamond^+$ is a PE operator.*

Proposition 6 gives us a constructive way to define an FE operator from a PE operator. Consider for instance the dilation operator $\diamond^+_{\mathsf{dil},d}$, where d is any distance between worlds (cf. Definition 5). Then the operator $\diamond_{\diamond^+_{\mathsf{dil},d}}$ consists in first "dilating" an input classifier φ so as to include all positive examples from μ^+, and then removing all instances introduced in the dilation step according to μ^-. As a consequence of Proposition 6, this operator satisfies **(F1-F5)**.

Likewise, each NE operator also defines an FE operator. Given an operator $\diamond^- : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$, $\diamond$ is said to be *negatively induced* by $\diamond^-$, denoted by $\diamond = \diamond_{\diamond^-}$, if it is defined for each classifier φ and each dataset (μ^+, μ^-) by $\varphi \diamond_{\diamond^-} (\mu^+, \mu^-) = (\varphi \diamond^- \mu^-) \vee \mu^+$. Echoing Proposition 6, we get that:

Proposition 7. *$\diamond_{\diamond^-}$ is an FE operator if and only if $\diamond^-$ is an NE operator.*

Remark that inducing an operator $\diamond$ by a PE operator *and* an NE operator, e.g., as $\varphi \diamond (\mu^+, \mu^-) = (\varphi \diamond^+ \mu^+) \diamond^- \mu^-$, does not always define an FE operator, even when $\diamond^+$ and $\diamond^-$ are dual. To give an example when this kind of construction does not work, let us consider our loan scenario again:

Example 1 (continued)**.** *Assume now that Alice receives an additional applicant, Emir, with profile $\omega_3 = 0100$. Since Emir has a low income, debts, and has not yet reimbursed his previous loan, Alice is certain that Emir is not eligible for a new loan. We are then given both a dataset $\mu = (\mu^+, \mu^-)$, where $\mu^+ = \neg p \wedge s \wedge (q \leftrightarrow \neg r)$ with $[\mu^+] = \{\omega_1, \omega_2\}$ (Bob / Cindy and Dan are positive examples), and $\mu^- = \neg p \wedge q \wedge \neg r \wedge \neg s$, i.e., $[\mu^-] = \{\omega_3\}$ (Emir is a negative example). Let us consider the operator $\diamond$ defined by $\varphi \diamond \mu = (\varphi \diamond^+_{\mathsf{dil},d_H} \mu^+) \diamond^-_{\mathsf{ero},d_H} \mu^-$, i.e., the classifier is first edited according to μ^+ using the Hamming-based dilation edit $\diamond^+_{\mathsf{dil},d_H}$, and is then edited again according to μ^- using the Hamming-based erosion edit $\diamond^-_{\mathsf{ero},d_H}$, that is, the negative edit operator dual to $\diamond^+_{\mathsf{dil},d_H}$. Recall first that $\varphi' = \varphi \diamond^+_{\mathsf{dil},d_H} \mu^+ = p \vee q \vee r$ (cf. Fig. 1c). Then we get that $\varphi'' = \varphi \diamond \mu = \varphi' \diamond^-_{\mathsf{ero},d_H} \mu^- \equiv (p \vee q) \wedge (r \vee (p \wedge q))$, with $[\varphi''] = D^{d_H}_{\varphi}(1)$ (cf. Fig. 1b). Yet $\{\omega_1, \omega_2\} \cap [\varphi''] = \emptyset$, i.e., Bob / Cindy and Dan are not classified as positive instances in the edited classifier φ''. Hence, $\diamond$ does not satisfy **(F1)**, i.e., $\diamond$ is not an FE operator.*

At that stage, a natural question is whether one can find an FE operator that is not induced by a PE operator or an NE operator. We provide below a positive answer to this question. In fact, we intend to introduce an operator which is not "decomposable" in any way by means of a combination of a PE operator and an NE operator. Formally, given an FE operator $\diamond$, a PE operator $\diamond^+$ and an NE operator $\diamond^-$, we say that the pair $(\diamond^+, \diamond^-)$ is *faithful to* $\diamond$ if for each classifier φ and each dataset (μ^+, μ^-), $\varphi \diamond (\mu^+, \mu^-) \equiv (\varphi \diamond^+ \mu^+) \diamond^- \mu^-$ or $\varphi \diamond (\mu^+, \mu^-) \equiv (\varphi \diamond^- \mu^-) \diamond^+ \mu^+$. An FE operator $\diamond$ is then said to be *decomposable* if $\diamond$ admits a faithful pair $(\diamond^+, \diamond^-)$. In particular, positively induced FE operators $\diamond_{\diamond^+}$ are decomposable: it can be easily verified that for each PE operator $\diamond^+$, the FE operator $\diamond_{\diamond^+}$ admits the faithful pair $(\diamond^+, \diamond^-_B)$, where $\diamond^-_B$ is the basic NE operator (recall that $\varphi \diamond^-_B \mu = \varphi \wedge \neg\mu$, for each classifier φ and each negative dataset μ). And similarly, negatively induced FE operators $\diamond_{\diamond^-}$ are decomposable as well since they admit the faithful pair $(\diamond^+_B, \diamond^-)$.

Now, let us consider the operator $\diamond_*$ defined for each classifier φ and each dataset (μ^+, μ^-) as:

$$\varphi \diamond_* (\mu^+, \mu^-) \begin{cases} \varphi, & \text{if } \mu^+ \models \varphi \text{ and } \mu^- \models \neg\varphi, \\ \top, & \text{if } \mu^+ \not\models \varphi \text{ and } [\mu^-] = \emptyset, \\ \mu^+, & \text{otherwise.} \end{cases}$$

This operator simply leaves unchanged the edited classifier in the case where it already complies with the input dataset (as required by **(F3)**). In the remaining cases, it behaves like the PE trivial operator $\diamond^+_T$ if the negative batch is empty, otherwise it behaves like the PE drastic operator $\diamond^+_D$ (cf. Definition 3). We can show that:

Proposition 8. *$\diamond_*$ is an FE operator that is not decomposable.*

This leaves us the interesting open question of whether a characterization result for FE operators can be found, that is left for further research.

Related Work

Theory revision is a change operation studied by ML researchers in the nineties that is connected to the edit one. Theory revision is an important component for concept formation, and as such it has been investigated and implemented as part of knowledge acquisition and machine learning systems (see e.g., MOBAL (Morik et al. 1994) and EITHER (Ourston and Mooney 1994)). Typically, in theory revision, a theory Σ is a logical representation (most of the time, a FOL formula) linking together atoms, denoting features used for describing instances and targeted concepts. An instance $\boldsymbol{x}$ is classified by Σ as an element of a concept y (represented by an atom) whenever y can be deduced from Σ and $\boldsymbol{x}$. When an instance together with its right concept (given by the change formula) is not classified by Σ as expected, a theory revision operator can be exploited to modify Σ so as to ensure that the instance is not classified incorrectly any longer in the revised theory. AGM contraction operators (Alchourrón, Gärdenfors, and Makinson 1985) can be used to this end (Wrobel 1993). Note that it can be the case that an instance $\boldsymbol{x}$ is not classified by Σ as an element of any concept. Accordingly, the representation to be changed Σ does not necessarily represent a "full" classifier as in the edit case. Furthermore, the basic operations that are used to derive the revised theory are usually not syntax-independent. This is typically the case, e.g., in the *learning from interpretations* setting (De Raedt 1997; De Raedt and Dehaspe 1997) where both the set of examples and the revised theory (called hypothesis) are full clausal theories, and other various formalizations of concept learning in logical settings, including inductive logic programming (Muggleton and De Raedt 1994; Flach 1997). This also makes them distinct from edit operations. Finally, works on theory revision are typically focused on defining specific approaches to achieve a revision of the input theory (possibly using few basic operations (Goldsmith et al. 2004; Goldsmith and Sloan 2005)), but they do not adopt an axiomatic perspective for delineating all the rational theory revision operators.

More recently, Zhou (2019) emphasized again the importance of integrating learning and reasoning in modern learning systems. The idea is to improve the decisions made by an underlying ML system C (the classifier), taking advantage of a reasoning module M. Roughly speaking, whenever a prediction P is made by the classifier C, it is transmitted to a reasoning module that checks whether the prediction is correct or not. If it is correct, nothing should be changed; otherwise, the corrected prediction P' found by M is transmitted back to C that is trained again using P'. Our edit framework is similar in essence, where a classifier φ plays the role of C and the positive / negative dataset μ is provided by an underlying expert module M. One of the strengths of Zhou's approach is that it is model-agnostic: the ML system can be any black box function. This is reminiscent to our edit framework where no further assumption is made on the representation of the input classifier φ and batch μ, besides being propositional formulae. However, in (Zhou 2019), the correction step is achieved by learning, which means that there is no guarantee that the repair is effective in the general case. In comparison, our framework, by its principled nature, guarantees the classifier to become fully compliant with the input batch after edit (cf. **(P1)**, **(N1)**, **(F1)** and **(F2)**).

Modifying a predictor so as to better take account for instances that are misclassified, as done with edit operations, is also at the core of *boosting*, a key principle in ML. In adaptative boosting for binary classification (Freund and Schapire 1997), the predictor has the form of an ensemble of weak learners, often decision trees reduced to decision stumps. The output of those weak learners is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is an iterative learning algorithm: at each iteration, the algorithm samples the training set, taking account for the distribution given by the weights associated with the instances (at start the uniform distribution is considered), then it looks for a weak classifier which minimizes the total weighted error, uses this to calculate the error rate and the weight of the weak classifier that has been generated, and finally update the weights of the instances so as to favor at the next step the selection of instances that have been misclassified by the generated weak learner. After a preset number of iterations, the algorithm stops. It turns out that the boosted classifier generated after an iteration may still misclassify the instances that were already misclassified by the boosted classifier before the iteration. Accordingly, the update operation at work in AdaBoost for improving the current boosted tree at each iteration is not a positive edit operator: **(P1)** is not satisfied.

Lastly, closely related to our work is a recent paper about *classifier rectification* (Coste-Marquis and Marquis 2021). Unlike the present paper, more than two classes can be considered in (Coste-Marquis and Marquis 2021) (classes are explicitly represented). Thus, two subsets X and Y of PS are used to encode, on the one hand, instances (positive ones and negative ones) and on the other hand, classes. When only two classes are targeted (the class of positive instances, a subset of Ω_X, the worlds over X, and its complementary set in Ω_X containing the negative instances), a singleton $Y = \{y\}$ is enough. Coste-Marquis and Marquis [2021] point out rules to be obeyed by any rational change operation $\star$ on Boolean classifiers Σ, when new pieces of evidence T must be taken into account. Boolean classifiers Σ are formulae from $\mathcal{L}$ satisfying the so-called XY-classification property. When $Y = \{y\}$, this precisely means that Σ is equivalent to $\varphi_X \Leftrightarrow y$ where φ_X is a formula over X. Thus, Σ classifies a given instance $\boldsymbol{x} \in \Omega_X$ as positive (resp. negative) whenever $\boldsymbol{x} \models \varphi_X$ (resp. $\boldsymbol{x} \models \neg\varphi_X$). Accordingly, every PE operation (resp. NE operation) of φ_X by a change formula μ_X corresponds to a rectification operation of $\Sigma = \varphi_X \Leftrightarrow y$ by $T = \mu_X \Rightarrow y$ (resp. $T = \mu_X \Rightarrow \neg y$). Postulates for the rectification operation have been provided in (Coste-Marquis and Marquis 2021). Though some connections between rectification postulates and PE / NE postulates exist, it is not the case that every PE (or NE) operator induces a rectification operator. Indeed, the rectification postulate **(RE2)** (see (Coste-Marquis and Marquis 2021) for details) makes formal a very demanding view of minimal change: when a change concerning an example (positive instance) μ_X is triggered, the classifications achieved by the rectified classifier coincide with those achieved by the classifier Σ at start, except possibly for μ_X (accordingly, rectification operators do not allow any generalization to take place and thus are not convenient for incremental learning).

Conclusion

The paper was focused on the question of editing Boolean classifiers, i.e., determining how a Boolean classifier should be modified when new pieces of evidence must be incorporated, an issue at the crossroads of ML and KR. Though the performance of ML models in terms of accuracy is impressive most of the time (especially when classifiers are learned from a sufficient amount of data), the error risk cannot be totally removed (this is intrinsic to inductive generalization). Thus, it is important to design and study approaches to determine how a classifier should be modified whenever it does not label an instance in the right way. The reported work is a step in this direction, centered on the identification of first principles (postulates) for characterizing what a rational change could be when dealing with Boolean classifiers.

One of our key assumptions in this paper is that the input dataset is fully reliable, which is reflected by the success postulate **(P1)** (in the case of positive edit). However, a number of standard learning algorithms take noisy examples into account, e.g., the k-NN algorithm, the perceptron algorithm, and algorithms for generating decision trees with pruning; and those algorithms do not satisfy **(P1)**. To extend the edit setting to noisy data, we plan to investigate how **(P1)** could be relaxed, so that an example is incorporated only if the corresponding piece of evidence is considered "sufficiently often" by the learning algorithm. For capturing such a behavior, *improvement* appears as a promising candidate (Konieczny, Medina Grespan, and Pino Pérez 2010), and determining the extent to which edit and improvement could be combined looks as a valuable perspective for further work.

Acknowledgments

This work has benefited from the support of the JSPS KAKENHI Grant Number JP21H04905 and JST CREST Grant Number JPMJCR22D3, Japan. Pierre Marquis has benefited from the support of the AI Chair EXPEKCTATION (ANR-19-CHIA-0005-01) of the French National Research Agency. He was also partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No952215.

References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50: 510–530.

Alchourrón, C. E.; and Makinson, D. 1985. On the logic of theory change: safe contraction. *Studia Logica*, 44(4): 405–422.

Audemard, G.; Bellart, S.; Bounia, L.; Koriche, F.; Lagniez, J.-M.; and Marquis, P. 2021. On the Computational Intelligibility of Boolean Classifiers. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning (KR'21)*, 74–86.

Audemard, G.; Koriche, F.; and Marquis, P. 2020. On Tractable XAI Queries based on Compiled Representations. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR'20)*, 838–849.

Aygün, O.; and Sönmez, T. 2013. Matching with Contracts: Comment. *American Economic Review*, 103(5): 2050–2051.

Bloch, I.; and Lang, J. 2002. *Towards mathematical morpho-logics*, 367–380. Heidelberg: Physica-Verlag HD.

Caridroit, T.; Konieczny, S.; and Marquis, P. 2017. Contraction in propositional logic. *International Journal of Approximate Reasoning*, 80: 428–442.

Coste-Marquis, S.; and Marquis, P. 2021. On Belief Change for Multi-Label Classifier Encodings. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, 1829–1836.

Dalal, M. 1988. Investigations into a theory of knowledge base revision: preliminary report. In *Proceedings of the 7th National Conference on Artificial Intelligence (AAAI'88)*, 475–479.

De Raedt, L. 1997. Logical Settings for Concept-Learning. *Artificial Intelligence*, 95(1): 187–201.

De Raedt, L.; and Dehaspe, L. 1997. Clausal Discovery. *Machine Learning*, 26(2-3): 99–146.

Domingos, P. M.; and Hulten, G. 2000. Mining high-speed data streams. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (SIGKDD'00)*, 71–80.

Fisher, D. H. 1987. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning*, 2(2): 139–172.

Flach, P. A. 1997. Normal Forms for Inductive Logic Programming. In *Proceedings of the 7th International Workshop of Inductive Logic Programming (ILP'97)*, 149–156.

Freund, Y.; and Schapire, R. E. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1): 119–139.

Gallant, S. I. 1988. Connectionist Expert Systems. *Communications of the ACM*, 31(2): 152–169.

Gärdenfors, P. 1988. *Knowledge in flux*. MIT Press.

Goldsmith, J.; and Sloan, R. H. 2005. New Horn Revision Algorithms. *Journal of Machine Learning Research*, 6: 1919–1938.

Goldsmith, J.; Sloan, R. H.; Szörényi, B.; and Turán, G. 2004. Theory revision with queries: Horn, read-once, and parity formulas. *Artificial Intelligence*, 156(2): 139–176.

Hatfield, J. W.; and Milgrom, P. R. 2005. Matching with Contracts. *American Economic Review*, 95(4): 913–935.

Ignatiev, A. 2020. Towards Trustable Explainable AI. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI'20)*, 5154–5158.

Ignatiev, A.; and Silva, J. M. 2021. SAT-Based Rigorous Explanations for Decision Lists. In *Proceedings of the 24th International Conference on Theory and Applications of Satisfiability Testing (SAT'21)*, 251–269.

Izza, Y.; Ignatiev, A.; and Marques-Silva, J. 2020. On Explaining Decision Trees. *CoRR*, abs/2010.11034.

Izza, Y.; and Marques-Silva, J. 2021. On Explaining Random Forests with SAT. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, 2584–2591.

Katsuno, H.; and Mendelzon, A. O. 1991. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52: 263–294.

Konieczny, S.; Medina Grespan, M.; and Pino Pérez, R. 2010. Taxonomy of Improvement Operators and the Problem of Minimal Change. In *Proceedings of the 12th International Conference on Principles of Knowledge Representation and Reasoning (KR'10)*, 161–170.

Maloof, M. A.; and Michalski, R. S. 2000. Selecting Examples for Partial Memory Learning. *Machine Learning*, 41(1): 27–52.

Maloof, M. A.; and Michalski, R. S. 2004. Incremental learning with partial instance memory. *Artificial Intelligence*, 154(1-2): 95–126.

Mitchell, T. M. 1977. Version Spaces: A Candidate Elimination Approach to Rule Learning. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI'77)*, 305–310.

Mitchell, T. M. 1982. Generalization as Search. *Artificial Intelligence*, 18(2): 203–226.

Morik, K.; Potamias, G.; Moustakis, V.; and Charissis, G. 1994. Knowledgeable learning using MOBAL: A medical case study. *Applied Artificial Intelligence*, 8(4): 579–592.

Muggleton, S.; and De Raedt, L. 1994. Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19/20: 629–679.

Narodytska, N.; Kasiviswanathan, S. P.; Ryzhyk, L.; Sagiv, M.; and Walsh, T. 2018. Verifying Properties of Binarized Deep Neural Networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 6615–6624.

Ourston, D.; and Mooney, R. 1994. Theory Refinement Combining Analytical and Empirical Methods. *Artificial Intelligence*, 66: 273–309.

Plotkin, G. D. 1970. A Note on Inductive Generalization. *Machine Intelligence*, 5: 153–163.

Quinlan, J. R. 1986. Induction of Decision Trees. *Machine Learning*, 1(1): 81–106.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6): 386–408.

Schlimmer, J. C.; and Fisher, D. H. 1986. A Case Study of Incremental Concept Induction. In *Proceedings of the 5th National Conference on Artificial Intelligence (AAAI'86)*, 496–501.

Shi, W.; Shih, A.; Darwiche, A.; and Choi, A. 2020. On Tractable Representations of Binary Neural Networks. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR'20)*, 882–892.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A Symbolic Approach to Explaining Bayesian Network Classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, 5103–5111.

Shih, A.; Choi, A.; and Darwiche, A. 2019. Compiling Bayesian Networks into Decision Graphs. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19)*, 7966–7974.

Utgoff, P. E. 1989. Incremental Induction of Decision Trees. *Machine Learning*, 4: 161–186.

Wrobel, S. 1993. On the Proper Definition of Minimality in Specialization and Theory Revision. In *Proceedings on the European Conference on Machine Learning (ECML'93)*, 65–82.

Yang, Y.-Y. 2020. Rationalizable choice functions. *Games and Economic Behavior*, 123: 120–126.

Zhou, Z. 2019. Abductive learning: towards bridging machine learning and logical reasoning. *Science China Information Science*, 62(7): 76101:1–76101:3.

Appendix: Proofs of Propositions

Proposition 2. An operator $\diamond^+$ is a PR operator if and only if there is a positive assignment $\varphi \mapsto f_\varphi$ such that for each classifier φ and each positive dataset μ,

$$[\varphi \diamond^+ \mu] = f_\varphi([\mu]).$$

Proof. (If part) From a great assignment $\varphi \mapsto f_\varphi$, define an operator $\diamond^+$ as $[\varphi \diamond^+ \mu] = f_\varphi([\mu])$, for each classifier φ and each positive dataset μ. We need to show that $\diamond^+$ satisfies **(P1-P4)**. **(P1)**, **(P2)** and **(P3)** are direct from conditions 1, 2, and 3 of a positive assignment, respectively. It remains to prove that $\diamond^+$ satisfies **(P4)**. Let φ be a classifier, μ be a positive dataset and ψ be a formula, and assume that $\psi \models \varphi \diamond^+ \mu$. We must show that $\varphi \diamond^+ \mu \equiv \varphi \diamond^+ (\mu \vee \psi)$. Let $W = [\mu]$ and $W' = [\mu \vee \psi]$. By definition of f_φ, we have that $f_\varphi(W) = [\varphi \diamond^+ \mu]$ and that $f_\varphi(W') = [\varphi \diamond^+ (\mu \vee \psi)]$. By condition 2 of a positive assignment, we know that $W \subseteq f_\varphi(W)$. Yet we assumed that $\psi \models \varphi \diamond^+ \mu$, that is, $[\psi] \subseteq f_\varphi(W)$. So by definition of W', i.e., since $W' = [\mu] \cup [\psi] = W \cup [\psi]$, we get that $W' \subseteq f_\varphi(W)$. We got that $W \subseteq f_\varphi(W)$ and $W' \subseteq f_\varphi(W)$, so by condition 4 of a positive assignment, we get that $f_\varphi(W) = f_\varphi(W')$, or stated otherwise, that $\varphi \diamond^+ \mu \equiv \varphi \diamond^+ (\mu \vee \psi)$. This shows that $\diamond^+$ satisfies **(P4)** and concludes the (if) part of the proof.

(Only if part) From a PE operator $\diamond^+$, associate every classifier φ with a mapping $f_\varphi : \mathcal{P}(\Omega) \mapsto \mathcal{P}(\Omega)$ defined for each set $W \subseteq \Omega$ as $f_\varphi(W) = [\varphi \diamond^+ \mu]$, where μ is some arbitrary chosen positive dataset such that $[\mu] = W$. We need to show that conditions 1-4 of a positive assignment are satisfied. Conditions 1, 2 and 3 are direct from **(P1)**, **(P2)**, and **(P3)**, respectively. To show that condtion 4 is satisfied, let $W, W' \subseteq \Omega$ such that $W \subseteq W'$ and $W' \subseteq f_\varphi(W)$. Consider now any arbitrary chosen positive dataset μ such that $[\mu] = W$, and let ψ be any formula such that $[\psi] = W' \setminus W$. Note that $W' = [\mu \vee \psi]$. Since $f_\varphi(W) = [\varphi \diamond^+ \mu]$ by definition, and since $W' \subseteq f_\varphi(W)$, we also get that $[\psi] \subseteq [\varphi \diamond^+ \mu]$, or equivalently, that $\psi \models \varphi \diamond^+ \mu$. Yet by **(P4)**, we get that $\varphi \diamond^+ \mu \equiv \varphi \diamond^+ (\mu \vee \psi)$, i.e., that $[\varphi \diamond^+ \mu] = [\varphi \diamond^+ (\mu \vee \psi)]$. By definition of $f_\varphi(\mu)$ and $f_\varphi(\mu \vee \psi)$, this means that $f_\varphi(\mu) = f_\varphi(\mu \vee \psi)$, that is, $f_\varphi(W) = f_\varphi(W')$. This shows that condition 4 is satisfied, and concludes the proof. □

Proposition 3. For each positive assignment $\varphi \mapsto f_\varphi$ and each $W \subseteq \Omega$, we have that $f_\varphi(f_\varphi(W)) = f_\varphi(W)$.

Proof. By condition 1, we know that $W \subseteq f_\varphi(W)$. Since we also trivially have that $f_\varphi(W) \subseteq f_\varphi(W)$, by condition 4 we get that $f_\varphi(f_\varphi(W)) = f_\varphi(W)$. □

Proposition 4. For every distance d, the operators $\diamond^+_{\mathsf{dil},d}$ and $\diamond^+_{\mathsf{gen},d}$ are PE operators.

Proof. Let d be any distance between worlds, and let us first show that $\diamond^+_{\mathsf{dil},d}$ satisfies **(P1-P4)**.

It is easy to see that the postulates **(P1)** and **(P3)** are directly satisfied by definition of $\diamond^+_{\mathsf{dil},d}$.

Let us prove **(P2)**. Let φ be a classifier, μ be a positive dataset, and assume $\mu \models \varphi$, and let us show that $\varphi \diamond^+_{\mathsf{dil},d} \mu \equiv \varphi$. By definition of $d(\omega, \varphi)$, we know that $d(\omega, \varphi) = 0$ if and only if $\omega \in [\varphi]$, for each world ω. Hence, $D_0^\varphi = [\varphi]$. Yet $\mu \models \varphi$, and so $[\mu] \subseteq D_0^\varphi$ and thus $[\varphi \diamond^+_{\mathsf{dil},d} \mu] = D_0^\varphi$, i.e., $\varphi \diamond^+_{\mathsf{dil},d} \mu \equiv \varphi$.

Let us now prove **(P4)**. So let φ be a classifier, μ and ψ be two positive datasets and assume that $\psi \models \varphi \diamond^+_{\mathsf{dil},d} \mu$. We need to prove that $\varphi \diamond^+_{\mathsf{dil},d} \mu \equiv \varphi \diamond^+_{\mathsf{dil},d} (\mu \vee \psi)$. Let $k_1 = \min(\{k \mid [\mu] \subseteq D^d_\varphi(k)\})$ and $k_2 = \min(\{k \mid [\mu \vee \psi] \subseteq D^d_\varphi(k)\})$. Let us prove that $k_1 = k_2$. For every $k \geq 0$, we trivially have that if $[\mu \vee \psi] \subseteq D^d_\varphi(k)$, then $[\mu] \subseteq D^d_\varphi(k)$. Hence, we get that $k_1 \leq k_2$. Let us prove that $k_2 \leq k_1$. By definition of k_1, we have that $[\varphi \diamond^{+}_{\mathsf{dil},d} \mu] = D^d_\varphi(k_1)$ and $[\mu] \subseteq D^d_\varphi(k_1)$. Yet we assumed that $\psi \models \varphi \diamond^{+}_{\mathsf{dil},d} \mu$, i.e., that $[\psi] \subseteq D^d_\varphi(k_1)$. Hence, we get that $[\mu \vee \psi] \subseteq D^d_\varphi(k_1)$, By definition of k_2, this means that $k_2 \leq k_1$. We got that $k_1 = k_2$, which shows by definition of $[\varphi \diamond^{+}_{\mathsf{dil},d} \mu]$ and $[\varphi \diamond^{+}_{\mathsf{dil},d} (\mu \vee \psi)]$ that $\varphi \diamond^{+}_{\mathsf{dil},d} \mu \equiv \varphi \diamond^{+}_{\mathsf{dil},d} (\mu \vee \psi)$. Hence, $\diamond^{+}_{\mathsf{dil},d}$ satisfies **(P4)**.

Let us now show that $\diamond^{+}_{\mathsf{gen},d}$ satisfies **(P1-P4)**.

It is easy to see that the postulate **(P3)** is directly satisfied by definition of $\diamond^{+}_{\mathsf{gen},d}$.

Let us prove **(P1)**. Let φ be a classifier and μ be a positive dataset. We must show that $\mu \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$. So let $\omega \in [\mu]$ and let us prove that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$. Let $\omega' \in \min([\varphi], \leq^d_\omega)$. We trivially have that $d(\omega, \omega) + d(\omega, \omega') \leq d(\omega, \omega')$. So by definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, we get that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$. We got that $\mu \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$, whish shows that $\diamond^{+}_{\mathsf{gen},d}$ satisfies **(P1)**.

Let us now prove **(P2)**. Let φ be a classifier, μ be a positive dataset, and assume that $\mu \models \varphi$. We must show that $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \equiv \varphi$. Let us first prove that $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \models \varphi$. So let $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, and let us show that $\omega \in [\varphi]$. By definition of $\varphi \diamond^{+}_{\mathsf{gen},d} \mu$, we know that there exist two worlds $\mu'\mu''$ such that $\omega' \in [\mu]$, $\omega'' \in \min([\varphi], \leq^d_{\omega'})$, and $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$. Since $\mu \models \varphi$ and $\omega' \in [\mu]$, we know that $\omega' \in [\varphi]$. So $\min([\varphi], \leq^d_{\omega'}) = \{\omega'\}$. This means that $\omega'' = \omega'$. So the equation $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$ can be equivalently written as $d(\omega, \omega') + d(\omega, \omega') \leq d(\omega', \omega')$, i.e., $d(\omega, \omega') = 0$. We got that $\omega = \omega'$. Yet we have seen that $\omega' \in [\varphi]$, thus $\omega \in [\varphi]$. This shows that $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \models \varphi$. Let us now prove that $\varphi \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$. So let $\omega \in [\varphi]$, and let us show that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$. Since $\omega \in [\varphi]$, we have that $\min([\varphi], \leq^d_\omega) = \{\omega\}$. Let ω' be any world such that $\omega' \in [\mu]$. By setting $\omega'' = \omega$, we get that $d(\omega, \omega') + d(\omega, \omega'') = d(\omega, \omega')$ on the one hand, and $d(\omega', \omega'') = d(\omega, \omega')$ on the other hand. So we get that $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$. So we have found two worlds ω', $\omega'' = \omega$, such that $\omega' \in [\mu]$, $\omega'' \in \min([\varphi], \leq^d_\omega)$, and $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$. By definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, this means that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$. This shows that $\varphi \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$. We got that $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \models \varphi$ and $\varphi \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$. Hence, $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \equiv \varphi$. This proves that $\diamond^{+}_{\mathsf{gen},d}$ satisfies **(P2)**.

Let us now prove **(P4)**. So let φ be a classifier, μ and ψ be two positive datasets and assume that $\psi \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$. We need to prove that $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \equiv \varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)$. Let us first prove that $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \models \varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)$. Let $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, and let us show that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)]$. By definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, we know that there exist two worlds ω', ω'' such that $\omega' \in [\mu]$, $\omega'' \in \min([\varphi], \leq^d_{\omega'})$, and $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$. Since $\omega' \in [\mu]$, we also trivially have that $\omega' \in [\mu \vee \psi]$. Then we found two worlds ω', ω'', such that $\omega' \in [\mu \vee \psi]$, $\omega'' \in \min([\varphi], \leq^d_{\omega'})$, and $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$. By definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)]$, this means that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)]$. Hence, $\varphi \diamond^{+}_{\mathsf{gen},d} \mu \models \varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)$. Let us now prove that $\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi) \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$. Let $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)]$, and let us show that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$. By definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)]$, we know that there exist two worlds ω', ω'' such that $\omega' \in [\mu \vee \psi]$, $\omega'' \in \min([\varphi], \leq^d_{\omega'})$, and $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$. We fall into two cases: $\omega' \in [\mu]$ or $\omega' \in [\psi]$. If the first case holds, i.e., $\omega' \in [\mu]$, then we found two worlds ω', ω'' such that $\omega' \in [\mu]$, $\omega'' \in \min([\varphi], \leq^d_{\omega'})$, and $d(\omega, \omega') + d(\omega, \omega'') \leq d(\omega', \omega'')$. By definition of $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, this means that $\omega \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, which was to be proved. Consider now the second case, i.e., $\omega' \in [\psi]$. Since we assumed that $\psi \models \varphi \diamond^{+}_{\mathsf{gen},d} \mu$, we get that $\omega' \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$. By definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, this means that there exist two worlds ω'_2, ω''_2, such that $\omega'_2 \in [\mu]$, $\omega''_2 \in \min([\varphi], \leq^d_{\omega'_2})$, and $d(\omega', \omega'_2) + d(\omega', \omega''_2) \leq d(\omega'_2, \omega''_2)$. Let us prove an intermediate statement, i.e., that $d(\omega'_2, \omega'') = d(\omega'_2, \omega''_2)$. Since $\omega''_2 \in \min([\varphi], \leq^d_{\omega'_2})$ and since $\omega'' \in [\varphi]$, we get that (i) $d(\omega'_2, \omega''_2) \leq d(\omega'_2, \omega'')$. Let us prove that $d(\omega'_2, \omega'') \leq d(\omega'_2, \omega''_2)$. Since d satisfies the triangular inequality property, we know that (ii) $d(\omega'_2, \omega'') \leq d(\omega'_2, \omega') + d(\omega', \omega'')$. Yet $\omega'' \in \min([\varphi], \leq^d_{\omega'})$ and $\omega''_2 \in [\varphi]$, so (iii) $d(\omega', \omega'') \leq d(\omega', \omega''_2)$. Hence from (iii), by adding $d(\omega'_2, \omega')$ on each side of the equation we get that (iv) $d(\omega'_2, \omega') + d(\omega', \omega'') \leq d(\omega'_2, \omega') + d(\omega', \omega''_2)$, and thus from (ii) we can write that (v) $d(\omega'_2, \omega'') \leq d(\omega'_2, \omega') + d(\omega', \omega''_2)$. Yet we know by definition of ω'_2 and ω''_2 (i.e., by definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$ and the fact that $\omega' \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$) that $d(\omega'_2, \omega') + d(\omega', \omega''_2) \leq d(\omega'_2, \omega''_2)$. Hence from (v), we got that (vi) $d(\omega'_2, \omega'') \leq d(\omega'_2, \omega''_2)$. We have shown that (i) $d(\omega'_2, \omega''_2) \leq d(\omega'_2, \omega'')$ and (vi) $d(\omega'_2, \omega'') \leq d(\omega'_2, \omega''_2)$. Therefore,

$$d(\omega'_2, \omega'') = d(\omega'_2, \omega''_2). \tag{1}$$

In particular, since $\omega''_2 \in \min([\varphi], \leq^d_{\omega'_2})$ and since $\omega'' \in [\varphi]$, from Equation 1 above we also get that

$$\omega'' \in \min([\varphi], \leq^d_{\omega'_2}). \tag{2}$$

We intend to prove now that $d(\omega'_2, \omega) + d(\omega, \omega'') \leq d(\omega'_2, \omega'')$. On the one hand, we know by definition of ω' and ω'' (i.e., by definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)]$) that (vii) $d(\omega', \omega) + d(\omega, \omega'') \leq d(\omega', \omega'')$. On the other hand, since d satisfies the triangular inequality property, we know that (viii) $d(\omega'_2, \omega) \leq d(\omega'_2, \omega') + d(\omega', \omega)$. From (viii) and by adding $d(\omega, \omega'')$ on each side of the equation, we can write that (ix) $d(\omega'_2, \omega) + d(\omega, \omega'') \leq d(\omega'_2, \omega') + d(\omega', \omega) + d(\omega, \omega'')$. Yet from (vii), $d(\omega'_2, \omega') + d(\omega', \omega) + d(\omega, \omega'') \leq d(\omega'_2, \omega') + d(\omega', \omega'')$, and thus from (ix) we get that (x) $d(\omega'_2, \omega) + d(\omega, \omega'') \leq d(\omega'_2, \omega') + d(\omega', \omega'')$. We know by definition of ω' and ω'' (i.e., by definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} (\mu \vee \psi)]$) that $\omega'' \in \min([\varphi], \leq^d_{\omega'})$. Since $\omega''_2 \in \varphi$, we get that $d(\omega', \omega'') \leq d(\omega', \omega''_2)$. So from (x), we can write that (xi) $d(\omega'_2, \omega) + d(\omega, \omega'') \leq d(\omega'_2, \omega') + d(\omega', \omega''_2)$. Yet since $\omega' \in [\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$, by definition of $[\varphi \diamond^{+}_{\mathsf{gen},d} \mu]$ we know that

$d(\omega_2', \omega') + d(\omega', \omega_2'') \le d(\omega_2', \omega_2'')$. So from (xi) we can write that (xii) $d(\omega_2', \omega) + d(\omega, \omega'') \le d(\omega_2', \omega_2'')$. And from (xii) and using Equation 1, we get that

$$d(\omega_2', \omega) + d(\omega, \omega'') \le d(\omega_2', \omega''). \qquad (3)$$

So we have found two worlds ω_2', ω'' such that $\omega_2' \in [\mu]$, $\omega'' \in \min([\varphi], \le^d_{\omega_2'})$ (cf. Equation 2), and $d(\omega, \omega_2') + d(\omega, \omega'') \le d(\omega_2', \omega'')$ (cf. Equation 3). This shows that $\omega \in [\varphi \diamond^+_{\mathsf{gen},d} \mu]$, which was to be proved. Indeed, we have shown that if $\omega \in [\varphi \diamond^+_{\mathsf{gen},d} (\mu \vee \psi)]$, then $\omega \in [\varphi \diamond^+_{\mathsf{gen},d} \mu]$. This proves that $\varphi \diamond^+_{\mathsf{gen},d} (\mu \vee \psi) \models \varphi \diamond^+_{\mathsf{gen},d} \mu$. Overall, we have shown that $\varphi \diamond^+_{\mathsf{gen},d} \mu \models \varphi \diamond^+_{\mathsf{gen},d} (\mu \vee \psi)$ and $\varphi \diamond^+_{\mathsf{gen},d} (\mu \vee \psi) \models \varphi \diamond^+_{\mathsf{gen},d} \mu$, thus $\varphi \diamond^+_{\mathsf{gen},d} \mu \equiv \varphi \diamond^+_{\mathsf{gen},d} (\mu \vee \psi)$. Hence, $\diamond^+_{\mathsf{gen},d}$ satisfies **(P4)**.

This concludes the proof. □

Proposition 5. σ is an involution, that is, for each operator $\diamond^* : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$, $\sigma(\sigma(\diamond^*)) = \diamond^*$. Moreover, $\sigma(\diamond^*)$ is an NE operator if and only if $\diamond^*$ is a PE operator.

Proof. Let us first show that σ is an involution. So let $\diamond^* : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$, and let us show that $\sigma(\sigma(\diamond^*)) = \diamond^*$. Let φ be any classifier and μ be any formula. Then:

$$\begin{aligned} \varphi\, \sigma(\sigma(\diamond^*))\, \mu &= \neg(\neg\varphi\, \sigma(\diamond^*)\, \mu) \\ &\equiv \neg(\neg(\neg\neg\varphi \diamond^* \mu)) \\ &\equiv \neg(\neg(\varphi \diamond^* \mu)) \\ &\equiv \varphi \diamond^* \mu. \end{aligned}$$

This shows that $\sigma(\sigma(\diamond^*)) = \diamond^*$, i.e., that σ is an involution.

Let us now show that $\sigma(\diamond^*)$ is an NE operator if and only if $\diamond^*$ is a PE operator.

(If part) Let $\diamond^+$ be a PE operator, and let $\diamond^- = \sigma(\diamond^+)$ and let us show that $\diamond^-$ is an NE operator, i.e., that $\diamond^-$ satisfies **(N1-N4)**.

(N1): let φ be any classifier and μ be any formula. We know that $\neg(\varphi \diamond^- \mu) \equiv \neg\varphi \diamond^+ \mu$. Yet $\diamond^+$ satisfies **(P1)**, so $\mu \models \neg\varphi \diamond^+ \mu$. So $\mu \models \neg(\varphi \diamond^- \mu)$. Hence, $\diamond^-$ satisfies **(N1)**.

(N2): let φ be any classifier, μ be any formula, and assume that $\mu \models \neg\varphi$. By **(P2)**, $\neg\varphi \diamond^+ \mu \equiv \neg\varphi$. Yet $\neg\varphi \diamond^+ \mu \equiv \neg(\varphi \diamond^- \mu)$. So $\neg(\varphi \diamond^- \mu) \equiv \neg\varphi$. Equivalently, $\varphi \diamond^- \mu \equiv \varphi$. Hence, $\diamond^-$ satisfies **(N2)**.

(N3): let φ_1, φ_2 be two classifiers, and μ_1, μ_2 be two formulae. Then $\varphi_1 \diamond^- \mu_1$ is equivalent to $\neg(\neg\varphi_1 \diamond^+ \mu_1)$ (by definition of $\diamond^- = \sigma(\diamond^+)$), which is equivalent to $\neg(\neg\varphi_2 \diamond^+ \mu_2)$ since $\diamond^+$ satisfies **(P3)**, which is equivalent to $\varphi_2 \diamond^- \mu_2$. Hence, $\diamond^-$ satisfies **(N3)**.

(N4): let φ be any classifier, μ and ψ be two formulae, and assume that $\psi \models \neg(\varphi \diamond^- \mu)$. So $\psi \models \neg\varphi \diamond^+ \mu$. Since $\diamond^+$ satisfies **(P4)**, we get that $\neg\varphi \diamond^+ \mu \equiv \neg\varphi \diamond^+ (\mu \vee \psi)$. So $\neg(\varphi \diamond^- \mu) \equiv \neg\varphi \diamond^+ \mu \equiv \neg\varphi \diamond^+ (\mu \vee \psi) \equiv \neg(\varphi \diamond^- (\mu \vee \psi))$. Hence, $\diamond^-$ satisfies **(N4)**.

(Only if part) Let $\diamond^+ : \mathcal{L} \times \mathcal{L} \mapsto \mathcal{L}$, let $\diamond^- = \sigma(\diamond^+)$ and assume that $\diamond^-$ is an NE operator. Remark that $\diamond^+ = \sigma(\diamond^-)$ since σ is an involution. We must prove that $\diamond^+$ is a PE operator, i.e., that $\diamond^+$ satisfies **(P1-P4)**. This is done in a very similar way as the previous (if part) of the proof, but we explicitely provide the proof below for the sake of completeness.

(P1): let φ be any classifier and μ be any formula. We know that $\varphi \diamond^+ \mu \equiv \neg(\neg\varphi \diamond^- \mu)$. Yet $\diamond^-$ satisfies **(N1)**, so $\mu \models \neg(\neg\varphi \diamond^- \mu)$. So $\mu \models \varphi \diamond^+ \mu$. Hence, $\diamond^+$ satisfies **(P1)**.

(P2): let φ be any classifier, μ be any formula, and assume that $\mu \models \varphi$. Equivalently, $\mu \models \neg\neg\varphi$. By **(N2)**, $\neg\varphi \diamond^- \mu \equiv \neg\varphi$. Yet $\neg\varphi \diamond^- \mu \equiv \neg(\varphi \diamond^+ \mu)$. So $\neg(\varphi \diamond^+ \mu) \equiv \neg\varphi$. Equivalently, $\varphi \diamond^+ \mu \equiv \varphi$. Hence, $\diamond^+$ satisfies **(P2)**.

(P3): let φ_1, φ_2 be two classifiers, and μ_1, μ_2 be two formulae. Then $\varphi_1 \diamond^+ \mu_1$ is equivalent to $\neg(\neg\varphi_1 \diamond^- \mu_1)$ (by definition of $\diamond^+ = \sigma(\diamond^-)$), which is equivalent to $\neg(\neg\varphi_2 \diamond^- \mu_2)$ since $\diamond^-$ satisfies **(N3)**, which is equivalent to $\varphi_2 \diamond^+ \mu_2$. Hence, $\diamond^+$ satisfies **(P3)**.

(P4): let φ be any classifier, μ and ψ be two formulae, and assume that $\psi \models \varphi \diamond^+ \mu$. So $\psi \models \neg(\neg\varphi \diamond^- \mu)$. Since $\diamond^-$ satisfies **(N4)**, we get that $\neg\varphi \diamond^- \mu \equiv \neg\varphi \diamond^- (\mu \vee \psi)$. So $\varphi \diamond^+ \mu \equiv \neg(\neg\varphi \diamond^- \mu) \equiv \neg(\neg\varphi \diamond^- (\mu \vee \psi)) \equiv \varphi \diamond^+ (\mu \vee \psi)$. Hence, $\diamond^+$ satisfies **(P4)**.

□

Proposition 6. $\diamond_{\diamond^+}$ is an FE operator if and only if $\diamond^+$ is a PE operator.

Proof. (If part) Let $\diamond^+$ be a PE operator, and let us prove that $\diamond_{\diamond^+}$ is an FE operator, i.e., that it satisfies **(F1-F5)**. Recall that $\diamond_{\diamond^+}$ is defined for each classifier φ and each dataset (μ^+, μ^-) by $\varphi \diamond_{\diamond^+} (\mu^+, \mu^-) = (\varphi \diamond^+ \mu^+) \wedge \neg\mu^-$.

(F1): since $\diamond^+$ satisfies **(P1)**, we know that $\mu^+ \models \varphi \diamond^+ \mu^+$. And since (μ^+, μ^-) is a dataset, by definition $[\mu^+] \wedge [\mu^-] = \emptyset$, or equivalently, $\mu^+ \models \neg\mu^-$. We got that $\mu^+ \models \varphi \diamond^+ \mu^+$ and $\mu^+ \models \neg\mu^-$, thus $\mu^+ \models (\varphi \diamond^+ \mu^+) \wedge \neg\mu^-$, i.e., $\mu^+ \models \varphi \diamond_{\diamond^+} (\mu^+, \mu^-)$. Hence, $\diamond_{\diamond^+}$ satisfies **(F1)**.

(F2): we trivially have that $\neg\mu^- \models (\varphi \diamond^+ \mu^+) \wedge \neg\mu^-$. So $\neg\mu^- \models \varphi \diamond_{\diamond^+} (\mu^+, \mu^-)$, or equivalently, $\mu^- \models \neg(\varphi \diamond_{\diamond^+} (\mu^+, \mu^-))$. Hence, $\diamond_{\diamond^+}$ satisfies **(F2)**.

(F3): assume that $\mu^+ \models \varphi$ and $\mu^- \models \neg\varphi$. Since $\diamond^+$ satisfies **(P2)**, we know that $\varphi \diamond^+ \mu^+ \equiv \varphi$. Thus $\varphi \diamond_{\diamond^+} (\mu^+, \mu^-) = (\varphi \diamond^+ \mu^+) \wedge \neg\mu^- \equiv \varphi \wedge \neg\mu^-$. Yet $\mu^- \models \neg\varphi$, i.e., $\neg\mu^- \models \varphi$, so $\varphi \wedge \neg\mu^- \equiv \varphi$. Thus $\varphi \diamond_{\diamond^+} (\mu^+, \mu^-) \equiv \varphi$. Hence, $\diamond_{\diamond^+}$ satisfies **(F3)**.

(F4): the fact that $\diamond_{\diamond^+}$ satisfies **(F4)** is obvious by definition and since $\diamond^+$ satisfies **(P3)**.

(F5): assume that $\psi \models \varphi \diamond (\mu^+, \mu^-)$ and $\alpha \models \neg(\varphi \diamond (\mu^+, \mu^-))$, or equivalently, that $\neg\alpha \models \varphi \diamond (\mu^+, \mu^-)$. We get that $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) = (\varphi \diamond^+ (\mu^+ \vee \psi)) \wedge (\neg\mu^- \vee \alpha)$ (by definition), which is equivalent to $(\varphi \diamond^+ \mu^+) \wedge (\neg\mu^- \vee \alpha)$ (since $\diamond^+$ satisfies **(P4)**), which is equivalent to $(\varphi \diamond^+ \mu^+) \wedge \neg\mu^- \wedge \neg\alpha$, which is equivalent to $(\varphi \diamond^+ \mu^+) \wedge \neg\mu^-$ (since $\neg\alpha \models \varphi \diamond (\mu^+, \mu^-)$), which is equal to $\varphi \diamond_{\diamond^+} (\mu^+, \mu^-)$ (by definition). We got that $\varphi \diamond_{\diamond^+} (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$. Hence, $\diamond_{\diamond^+}$ satisfies **(F5)**.

We have shown that if $\diamond^+$ is a PE operator, then $\diamond_{\diamond^+}$ is an FE operator, which concludes the (if) part of the proof.

(Only if part) Let us first remark that by setting $\mu^- = \bot$, we get that $\varphi \diamond_{\diamond^+} (\mu^+, \bot) \equiv \varphi \diamond^+ \mu^+$. Now, assume that $\diamond_{\diamond^+}$ is an FE operator, and let us prove that $\diamond^+$ is an FE operator, i.e., it necessarily satisfies **(P1-P4)**.

(P1): since $\diamond_{\diamond^+}$ satisfies **(F1)**, $\mu^+ \models \varphi \diamond_{\diamond^+} (\mu^+, \bot)$. And since $\varphi \diamond^+ \mu^+ \equiv \varphi \diamond_{\diamond^+} (\mu^+, \bot)$, we get that $\mu^+ \models \varphi \diamond^+ \mu^+$. Hence, $\diamond^+$ satisfies **(P1)**.

(P2): assume that $\mu^+ \models \varphi$. We trivially have that $\bot \models \neg\varphi$. So since $\diamond_{\diamond^+}$ satisfies **(F3)**, we get that $\varphi \diamond_{\diamond^+} (\mu^+, \bot) \equiv \varphi$. Yet $\varphi \diamond^+ \mu^+ \equiv \varphi \diamond_{\diamond^+} (\mu^+, \bot)$, so $\varphi \diamond^+ \mu^+ \equiv \varphi$. Hence, $\diamond^+$ satisfies **(P2)**.

(P3): the fact that $\diamond^+$ satisfies **(P3)** is direct by definition and since $\varphi \diamond^+ \mu^+ \equiv \varphi \diamond_{\diamond^+} (\mu^+, \bot)$ and $\diamond_{\diamond^+}$ satisfies **(F4)**.

(P4): assume that $\psi \models \varphi \diamond^+ \mu^+$. Since $\varphi \diamond^+ \mu^+ \equiv \varphi \diamond_{\diamond^+} (\mu^+, \bot)$, we get that $\psi \models \varphi \diamond_{\diamond^+} (\mu^+, \bot)$. And by setting $\alpha = \bot$, we trivially have that $\alpha \models \varphi \diamond_{\diamond^+} (\mu^+, \bot)$. Since $\diamond_{\diamond^+}$ satisfies **(F5)**, we get that $\varphi \diamond_{\diamond^+} (\mu^+, \bot) \equiv \varphi \diamond_{\diamond^+} (\mu^+ \vee \psi, \bot \vee \alpha)$, i.e., $\varphi \diamond_{\diamond^+} (\mu^+, \bot) \equiv \varphi \diamond_{\diamond^+} (\mu^+ \vee \psi, \bot)$. Thus, $\varphi \diamond^+ \mu^+ \equiv \varphi \diamond^+ (\mu^+ \vee \psi)$. Hence, $\diamond^+$ satisfies **(P4)**.

We have shown that if $\diamond_{\diamond^+}$ is an FE operator, then $\diamond^+$ is a PE operator, which concludes the (only if) part of the proof.

This concludes the proof of the proposition. $\square$

Proposition 7. $\diamond_{\diamond^-}$ is an FE operator if and only if $\diamond^-$ is an NE operator.

Proof. The proof is very similar to the proof of Proposition 6, but we explicitely provide it for the sake of completeness.

(If part) Let $\diamond^-$ be an NE operator, and let us prove that $\diamond_{\diamond^-}$ is an FE operator, i.e., that it satisfies **(F1-F5)**. Recall that $\diamond_{\diamond^-}$ is defined for each classifier φ and each dataset (μ^+, μ^-) by $\varphi \diamond_{\diamond^-} (\mu^+, \mu^-) = (\varphi \diamond^- \mu^-) \vee \mu^+$.

(F1): we trivially have that $\mu^+ \models (\varphi \diamond^- \mu^-) \vee \mu^+$. So $\mu^+ \models \varphi \diamond_{\diamond^-} (\mu^+, \mu^-)$. Hence, $\diamond_{\diamond^+}$ satisfies **(F1)**.

(F2): since $\diamond^-$ satisfies **(N1)**, we know that $\mu^- \models \neg(\varphi \diamond^- \mu^-)$. And since (μ^+, μ^-) is a dataset, by definition $[\mu^+] \wedge [\mu^-] = \emptyset$, or equivalently, $\mu^- \models \neg\mu^+$. We got that $\mu^- \models \neg(\varphi \diamond^- \mu^-)$ and $\mu^- \models \neg\mu^+$, thus $\mu^- \models \neg(\varphi \diamond^- \mu^-) \wedge \neg\mu^+$, i.e., $\mu^- \models \neg((\varphi \diamond^- \mu^-) \vee \mu^+)$, that is, $\mu^- \models \neg(\varphi \diamond_{\diamond^-} (\mu^+, \mu^-))$. Hence, $\diamond_{\diamond^-}$ satisfies **(F2)**.

(F3): assume that $\mu^+ \models \varphi$ and $\mu^- \models \neg\varphi$. Since $\diamond^-$ satisfies **(N2)**, we know that $\varphi \diamond^- \mu^- \equiv \varphi$. Thus $\varphi \diamond_{\diamond^-} (\mu^+, \mu^-) = (\varphi \diamond^- \mu^-) \vee \mu^+ \equiv \varphi \vee \mu^+$. Yet $\mu^+ \models \varphi$, so $\varphi \vee \mu^+ \equiv \varphi$. Thus $\varphi \diamond_{\diamond^-} (\mu^+, \mu^-) \equiv \varphi$. Hence, $\diamond_{\diamond^-}$ satisfies **(F3)**.

(F4): the fact that $\diamond_{\diamond^-}$ satisfies **(F4)** is obvious by definition and since $\diamond^-$ satisfies **(N3)**.

(F5): assume that $\psi \models \varphi \diamond (\mu^+, \mu^-)$ and $\alpha \models \neg(\varphi \diamond (\mu^+, \mu^-))$. We get that $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) = (\varphi \diamond^- (\mu^- \vee \alpha)) \vee (\mu^+ \vee \psi)$ (by definition), which is equivalent to $(\varphi \diamond^- \mu^-) \vee (\mu^+ \vee \psi)$ (since $\diamond^-$ satisfies **(N4)**), which is equivalent to $(\varphi \diamond^- \mu^-) \vee \mu^+$ (since $\psi \models \varphi \diamond (\mu^+, \mu^-)$), which is equal to $\varphi \diamond_{\diamond^-} (\mu^+, \mu^-)$ (by definition). We got that $\varphi \diamond_{\diamond^-} (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$. Hence, $\diamond_{\diamond^-}$ satisfies **(F5)**.

We have shown that if $\diamond^-$ is an NE operator, then $\diamond_{\diamond^-}$ is an FE operator, which concludes the (if) part of the proof.

(Only if part) Let us first remark that by setting $\mu^+ = \bot$, we get that $\varphi \diamond_{\diamond^-} (\bot, \mu^-) \equiv \varphi \diamond^- \mu^-$. Now, assume that $\diamond_{\diamond^-}$ is an FE operator, and let us prove that $\diamond^-$ is an NE operator, i.e., it necessarily satisfies **(N1-N4)**.

(N1): since $\diamond_{\diamond^-}$ satisfies **(F1)**, $\mu^- \models \neg(\varphi \diamond_{\diamond^-} (\bot, \mu^-))$. And since $\varphi \diamond^- \mu^- \equiv \varphi \diamond_{\diamond^-} (\bot, \mu^-)$, we get that $\mu^- \models \neg(\varphi \diamond^- \mu^-)$. Hence, $\diamond^-$ satisfies **(N1)**.

(N2): assume that $\mu^- \models \neg\varphi$. We trivially have that $\bot \models \varphi$. So since $\diamond_{\diamond^-}$ satisfies **(F3)**, we get that $\varphi \diamond_{\diamond^-} (\bot, \mu^-) \equiv \varphi$. Yet $\varphi \diamond^- \mu^- \equiv \varphi \diamond_{\diamond^-} (\bot, \mu^-)$, so $\varphi \diamond^- \mu^- \equiv \varphi$. Hence, $\diamond^-$ satisfies **(N2)**.

(N3): the fact that $\diamond^-$ satisfies **(N3)** is direct by definition and since $\varphi \diamond^- \mu^- \equiv \varphi \diamond_{\diamond^-} (\bot, \mu^-)$ and $\diamond_{\diamond^-}$ satisfies **(F4)**.

(N4): assume that $\alpha \models \neg(\varphi \diamond^- \mu^-)$. Since $\varphi \diamond^- \mu^- \equiv \varphi \diamond_{\diamond^-} (\bot, \mu^-)$, we get that $\alpha \models \neg(\varphi \diamond_{\diamond^-} (\bot, \mu^-))$. And by setting $\psi = \bot$, we trivially have that $\psi \models \varphi \diamond_{\diamond^-} (\bot, \mu^-)$. Since $\diamond_{\diamond^-}$ satisfies **(F5)**, we get that $\varphi \diamond_{\diamond^-} (\bot, \mu^-) \equiv \varphi \diamond_{\diamond^-} (\bot \vee \psi, \mu^- \vee \alpha)$, i.e., $\varphi \diamond_{\diamond^-} (\bot, \mu^-) \equiv \varphi \diamond_{\diamond^-} (\bot, \mu^- \vee \alpha)$. Thus, $\varphi \diamond^- \mu^- \equiv \varphi \diamond^- (\mu^- \vee \alpha)$. Hence, $\diamond^-$ satisfies **(N4)**.

We have shown that if $\diamond_{\diamond^-}$ is an FE operator, then $\diamond^-$ is an NE operator, which concludes the (only if) part of the proof.

This concludes the proof of the proposition. $\square$

Proposition 8. $\diamond_*$ is an FE operator that is not decomposable.

The proof of Proposition 8 is based on the following lemma:

Lemma 1. *Every decomposable FE operator $\diamond$ admits a unique faithful pair $(\diamond^+, \diamond^-)$, defined for each classifier φ and each dataset (μ^+, μ^-) as*

$$\begin{cases} \varphi \diamond^+ \mu^+ \equiv \varphi \diamond (\mu^+, \bot), \text{ and} \\ \varphi \diamond^- \mu^- \equiv \varphi \diamond (\bot, \mu^-). \end{cases}$$

Proof. Let $\diamond$ be a decomposable FE operator and let $(\diamond^+, \diamond^-)$ be a pair that is faithful to $\diamond$. Let φ be a classifier and μ^+ be a positive dataset. We fall into two cases: (i) $\varphi \diamond (\mu^+, \bot) \equiv (\varphi \diamond^+ \mu^+) \diamond^- \bot$, or (ii) $\varphi \diamond (\mu^+, \bot) \equiv (\varphi \diamond^- \bot) \diamond^+ \mu^+$. In case (i), since $\diamond^-$ satisfies **(N2)** and $\bot \models \neg\varphi \diamond^+ \mu^+$, we get that $(\varphi \diamond^+ \mu^+) \diamond^- \bot \equiv \varphi \diamond^+ \mu^+$, hence $\varphi \diamond (\mu^+, \bot) \equiv \varphi \diamond^+ \mu^+$. In case (ii), since $\diamond^-$ satisfies **(N2)** and $\bot \models \neg\varphi$, we get that $\varphi \diamond^- \bot \equiv \varphi$, thus $(\varphi \diamond^- \bot) \diamond^+ \mu^+ \equiv \varphi \diamond^+ \mu^+$, and so $\varphi \diamond (\mu^+, \bot) \equiv \varphi \diamond^+ \mu^+$. From both cases (i) and (ii), we got for each classifier φ and each positive dataset μ^+ that $\varphi \diamond^+ \mu^+ \equiv \varphi \diamond (\mu^+, \bot)$.

Likewise, to prove that $\varphi \diamond^- \mu^- \equiv \varphi \diamond (\bot, \mu^-)$ for each classifier φ and each negative dataset μ^-, it is enough to see that since $\diamond^+$ satisfies **(P2)**, whether $\varphi \diamond (\bot, \mu^-) \equiv (\varphi \diamond^+ \bot) \diamond^- \mu^-$ or $\varphi \diamond (\bot, \mu^-) \equiv (\varphi \diamond^- \mu^-) \diamond^+ \bot$, in both cases we get that $\varphi \diamond^- \mu^- \equiv \varphi \diamond (\bot, \mu^-)$.

This concludes the proof of the lemma. $\square$

We are now ready to provide the proof of Proposition 8:

Proof. Let us first prove that $\diamond_*$ satisfies **(F1-F5)**. The postulates **(F3)** and **(F4)** are trivially satisfied by definition of $\diamond_*$. Let us prove that $\diamond_*$ satisfies **(F1)**, **(F2)** and **(F5)**. Let φ be any classifier and (μ^+, μ^-) be any dataset. In the proof below, let us consider and three separate cases: (i) $\mu^+ \models \varphi$

and $\mu^- \models \neg\varphi$; (ii) $\mu^+ \not\models \varphi$ and $[\mu^-] = \emptyset$; and (iii) the remaining cases. Obvously enough, we fall into one of these three cases.

(F1): we must prove that $\mu^+ \models \varphi \diamond_* (\mu^+, \mu^-)$. In case (i), $\varphi \diamond_* (\mu^+, \mu^-) = \varphi$. So we trivially get that $\mu^+ \models \varphi \diamond_* (\mu^+, \mu^-)$. In case (ii), we get that $\varphi \diamond_* (\mu^+, \mu^-) = \top$ and in case (iii), $\varphi \diamond_* (\mu^+, \mu^-) = \mu^+$, so we can see in all cases that $\mu^+ \models \varphi \diamond_* (\mu^+, \mu^-)$. Hence, $\diamond_*$ satisfies **(F1)**.

(F2): we must prove that $\mu^- \models \neg(\varphi \diamond_* (\mu^+, \mu^-))$. In case (i), $\varphi \diamond_* (\mu^+, \mu^-) = \varphi$. So we trivially get that $\mu^- \models \neg\varphi \diamond_* (\mu^+, \mu^-)$. In case (ii), since $[\mu^-] = \emptyset$, we trivially get that $\mu^- \models \neg(\varphi \diamond_* (\mu^+, \mu^-))$. And in case (iii), $\varphi \diamond_* (\mu^+, \mu^-) = \mu^+$. Yet by definition of a dataset, $[\mu^+] \cap [\mu^-] = \emptyset$. Thus $[\varphi \diamond_* (\mu^+, \mu^-)] \cap [\mu^-] = \emptyset$, or stated otherwise, $\mu^- \models \neg(\varphi \diamond_* (\mu^+, \mu^-))$. In all three cases we got that $\mu^- \models \neg(\varphi \diamond_* (\mu^+, \mu^-))$. Hence, $\diamond_*$ satisfies **(F2)**.

(F5): let ψ, α be two formulae such that $\psi \models \varphi \diamond_* (\mu^+, \mu^-)$ and $\alpha \models \neg(\varphi \diamond_* (\mu^+, \mu^-))$. We must prove that $\varphi \diamond (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$. Assume first that we fall into case (i), i.e., $\mu^+ \models \varphi$ and $\mu^- \models \neg\varphi$. Then $\varphi \diamond_* (\mu^+, \mu^-) = \varphi$. This means that $\psi \models \varphi$ and $\alpha \models \neg\varphi$. We got that $\mu^+ \vee \psi \models \varphi$ and $\mu^- \vee \alpha \models \neg\varphi$. So by definition of $\diamond_*$, $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) \equiv \varphi$. That is, $\varphi \diamond (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$. Assume now that we fall into case (ii), i.e., $\mu^+ \not\models \varphi$ and $[\mu^-] = \emptyset$. Then $\varphi \diamond_* (\mu^+, \mu^-) = \top$. In particular, since $\mu^+ \not\models \varphi$ we know that $\mu^+ \vee \psi \not\models \varphi$. And since $\alpha \models \neg(\varphi \diamond_* (\mu^+, \mu^-))$, this means that $[\alpha] = \emptyset$, thus $[\mu^- \vee \alpha] = \emptyset$. since $\mu^+ \not\models \varphi$ we know that $\mu^+ \vee \psi \not\models \varphi$. So we got that $\mu^+ \vee \psi \not\models \varphi$ and $[\mu^- \vee \alpha] = \emptyset$, which by definition of $\diamond_*$ means that $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) = \top$. We got that $\varphi \diamond (\mu^+, \mu^-) = \top$ and $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) = \top$, thus $\varphi \diamond (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$. Assume now that we fall into the remaining case (iii). It is easy to see that since this case corresponds to the negation of the disjunction of cases (i) and (ii), we fall into one of the following two subcases: either (iii-a) $\mu^- \not\models \neg\varphi$, or (iii-b) $\mu^+ \not\models \varphi$, $\mu^- \models \neg\varphi$, and $[\mu^-] \neq \emptyset$. And in this case, $\varphi \diamond (\mu^+, \mu^-) = \mu^+$. Now, in case (iii-a), we obviously get that $\mu^- \vee \alpha \not\models \varphi$, and thus $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) = \mu^+$ by definition. And in case (iii-b), we get that $\mu^+ \vee \psi \not\models \varphi$ and $[\mu^- \vee \alpha] \neq \emptyset$, and thus $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) = \mu^+$. So since $\varphi \diamond (\mu^+, \mu^-) = \mu^+$ and since in both cases (iii-a) and (iii-b), $\varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha) = \mu^+$, we got that $\varphi \diamond (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$. We have shown that in every case ((i), (ii), (iii-a) and (iii-b)), $\varphi \diamond (\mu^+, \mu^-) \equiv \varphi \diamond (\mu^+ \vee \psi, \mu^- \vee \alpha)$. Hence, $\diamond_*$ satisfies **(F5)**.

This concludes the proof that $\diamond_*$ satisfies **(F1-F5)**, i.e., $\diamond_*$ is an FE operator.

Let us now show that $\diamond_*$ is not decomposable. Assume toward a contradiction that $\diamond_*$ is decomposable. Then from Lemma 1, we know that $\diamond_*$ admits a unique faithful pair $(\diamond_*^+, \diamond_*^-)$ where $\diamond_*^+$ and $\diamond_*^-$ are defined that for each classifier φ and each dataset (μ^+, μ^-) as $\varphi \diamond_*^+ \mu^+ \equiv \varphi \diamond (\mu^+, \bot)$ and $\varphi \diamond_*^- \mu^- \equiv \varphi \diamond_* (\bot, \mu^-)$. Let φ be a classifier, (μ^+, μ^-) be a dataset, and assume that $[\varphi] = \{\omega^\varphi\}$, $[\mu^+] = \{\omega^+\}$ and $[\mu^-] = \{\omega^-\}$, where ω^φ, ω^+ and ω^- are three distinct worlds. Since $\diamond_*$ is decomposable, we know that $\varphi \diamond_* (\mu^+, \mu^-) \equiv (\varphi \diamond_*^+ \mu^+) \diamond_*^- \mu^-$ or $\varphi \diamond_* (\mu^+, \mu^-) \equiv (\varphi \diamond_*^- \mu^-) \diamond_*^+ \mu^+$. Let us show that both cases lead to a contradiction:

- Case 1: $\varphi \diamond_* (\mu^+, \mu^-) \equiv (\varphi \diamond_*^- \mu^-) \diamond_*^+ \mu^+$. Since $\varphi \diamond_*^- \mu^- \equiv \varphi \diamond_* (\bot, \mu^-)$, and by definition $\varphi \diamond_* (\bot, \mu^-) = \varphi$ since $\bot \models \varphi$ and $\mu^- \models \neg\varphi$, we get that $\varphi \diamond_*^- \mu^- \equiv \varphi$. Thus $\varphi \diamond_* (\mu^+, \mu^-) \equiv \varphi \diamond_*^+ \mu^+$. Yet $\varphi \diamond_*^+ \mu^+ \equiv \varphi \diamond_* (\mu^+, \bot)$, and by definition, $\varphi \diamond_* (\mu^+, \bot) = \top$ since $\mu^+ \not\models \varphi$ and $[\bot] = \emptyset$. So $\varphi \diamond_* (\mu^+, \mu^-) \equiv \top$. We got that $\mu^- \not\models \neg(\varphi \diamond_* (\mu^+, \mu^-))$, which contradicts the fact that $\diamond_*$ satisfies **(F2)**.
- Case 2: $\varphi \diamond_* (\mu^+, \mu^-) \equiv (\varphi \diamond_*^+ \mu^+) \diamond_*^- \mu^-$. On the one hand, we know by definition of $\diamond_*$ that (i) $\varphi \diamond_* (\mu^+, \mu^-) \equiv \mu^+$ since $\mu^+ \not\models \varphi$ and $[\mu^-] \neq \emptyset$. Yet $\varphi \diamond_*^+ \mu^+ \equiv \varphi \diamond_* (\mu^+, \bot)$, and by definition, $\varphi \diamond_* (\mu^+, \bot) = \top$ since $\mu^+ \not\models \varphi$ and $[\bot] = \emptyset$. So $\varphi \diamond_* (\mu^+, \mu^-) \equiv \top \diamond_*^- \mu^-$. Yet $\top \diamond_*^- \mu^- \equiv \top \diamond_* (\bot, \mu^-) \equiv \bot$ by definition of $\diamond_*$ and since $\mu^- \not\models \neg\top$ and $[\mu^-] \neq \emptyset$. Thus we got that $\varphi \diamond_* (\mu^+, \mu^-) \equiv \bot$. This contradicts (i), i.e., the fact that $\varphi \diamond_* (\mu^+, \mu^-) \equiv \mu^+$, since $[\mu^+] \neq \emptyset$.

We showed that assuming that $\diamond_*$ is decomposable leads to a contradiction. Therefore, $\diamond_*$ is not decomposable.

This concludes the proof. □