Iterated Belief Change as Learning (Extended Abstract)

Nicolas Schwind[1] , **Katsumi Inoue**[2] , **Sébastien Konieczny**[3] and **Pierre Marquis**[3,4]
[1]National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
[2]National Institute of Informatics, Tokyo, Japan
[3]Univ. Artois, CNRS, CRIL, Lens, France
[4]Institut Universitaire de France
nicolas-schwind@aist.go.jp, inoue@nii.ac.jp, konieczny@cril.fr, marquis@cril.fr

Abstract

In this work, we show how the class of improvement operators – a general class of iterated belief change operators – can be used to define a learning model. Focusing on binary classification, we present learning and inference algorithms suited to this learning model and we evaluate them empirically. Our findings highlight two key insights: first, that iterated belief change can be viewed as an effective form of online learning, and second, that the well-established axiomatic foundations of belief change operators offer a promising avenue for the axiomatic study of classification tasks. This paper is a summary of (Schwind et al. 2025).

1 Introduction

Belief Change Theory (BCT) (Alchourrón, Gärdenfors, and Makinson 1985; Katsuno and Mendelzon 1991) provides a principled framework for modifying an agent's beliefs in response to new information. Iterated belief revision (Darwiche and Pearl 1997) extends this framework to handle sequences of revisions. In both cases, the objective is to improve the agent's beliefs so they better reflect the world. While the methodologies differ, this goal aligns with that of Machine Learning (ML): deriving an accurate approximation of the world from data.

Despite this conceptual similarity, connections between BCT and ML remain largely unexplored. A major difference is that primacy of update (new information must be fully adopted) is central in belief revision. This is incompatible with typical ML settings involving noisy data, as it leads to drastic changes at each learning step.

Improvement operators (Konieczny and Pino Pérez 2008) generalize iterated belief revision by weaking the primacy of update principle. They allow for incremental changes, better capturing the gradual nature of learning. When the same input is received again, its plausibility is slightly adjusted. Formally, an improvement operator $\circ$ is defined over an *epistemic space* $\langle E, Bel\rangle$, where E is a set of *epistemic states* and Bel maps each state $\Psi \in E$ to a propositional formula. Each state $\Psi \in E$ is an object that encodes the agent's actual beliefs (extracted through the formula $Bel(\Psi)$) and conditional information guiding further revision steps. Then, an improvement operator $\circ$ maps a pair (Ψ, φ) (a state and a formula representing new input) to a modified state $\Psi \circ \varphi$ in which the plausibility of φ's models is slightly increased.

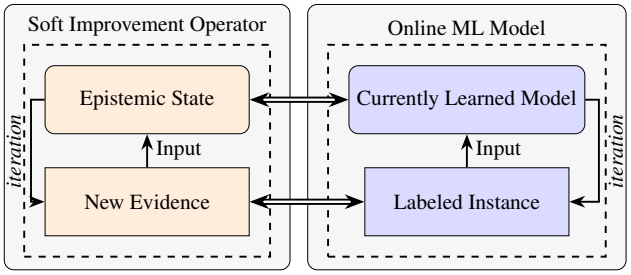

Figure 1: Analogies between classifier learning models and belief change operators.

A prominent example of epistemic space is $\langle E_{ocf}, Bel_{ocf}\rangle$, where E_{ocf} is the set of Ordinal Conditional Functions (OCFs) κ, which assign ranks (non-negative integers) to worlds. The lower the rank, the higher the plausibility. The agent's beliefs is a formula $Bel_{ocf}(\kappa)$ whose models are worlds ω such that $\kappa(\omega) = 0$. A simple improvement operator, the *basic shifting operator* $\circ_{+1}$, uniformly increases in the current epistemic state the plausibility (i.e., decreases the rank) of all φ-worlds by 1, where φ is the new input.

Improvement operators mirror online classification learning, where each received labeled example leads to gradual changes in its estimated likelihood of the underlying class membership. This analogy is illustrated in Figure 1.

We developed a learning framework inspired by this belief change perspective. We focused on binary classification and showed that models based on improvement operators yield reasonable performance, offering a promising new approach to learning from examples.

2 Proposal

From epistemic spaces to classifiers. We formalize binary classification in a propositional framework. Each instance is described by binary features $P_{\mathbf{X}} = \{x_1, \ldots, x_n\}$, and the task is to predict whether it belongs to the target class (output 1 or 0). A *labeled* instance extends the feature set with a class variable $\mathbf{y}$, forming a world over $P = P_{\mathbf{X}} \cup \mathbf{y}$. Labeled instances ω are represented as complete formulas φ_ω, and a training dataset D is a sequence of such formulas $(\varphi^s_\omega)_{1\leq s\leq m}$. The set of all possible datasets is denoted $\mathcal{D}$.

Following the formalization of improvement operators on epistemic spaces, we introduce a corresponding notion of *classifier space* $\langle E, Pos\rangle$, where E is a set of classifiers and Pos maps each classifier to a propositional formula predicting the target class.

A learning operator $\odot$ defines how a classifier is revised upon receiving a labeled example. Combined with an *anchor* $\Psi_* \in E$ (an initial classifier), this yields a learning framework $(\Psi_*, \odot)$, modeling classifier evolution through sequential revisions: $\Psi_* \odot \emptyset = \Psi_*$, and $\Psi_* \odot (\mathsf{D} \sqcup (\varphi_\omega)) = (\Psi_* \odot \mathsf{D}) \odot \varphi_\omega$. A classifier thus evolves through training, just as beliefs evolve through iterated revision.

Improvement-based learning. We define a family of learning operators over classifiers represented as tuples $\Psi = (\mathsf{D}, \kappa, \tau)$, called TOCFS. Here, D is a training dataset, κ an OCF over instances, and $\tau \in \mathbb{R}$ a threshold: an instance $\omega_{\mathbf{X}}$ is classified as positive iff $\kappa(\omega_{\mathbf{X}}) \leq \tau$. This classifier space is denoted $\langle E_{tocf}, Pos_{tocf}\rangle$, where $E_{tocf} = \mathcal{D} \times E_{ocf} \times \mathbb{R}$ and Pos_{tocf} maps each TOCF to the formula satisfied by all instances $\omega_{\mathbf{X}}$ with $\kappa(\omega_{\mathbf{X}}) \leq \tau$.

Each operator $\odot_{(\circ, B, \mathbf{m})}$ in this family is defined on $\langle E_{tocf}, Pos_{tocf}\rangle$ and characterized by (i) an improvement operator $\circ$ on OCFs, (ii) a neighborhood B over instances, and (iii) a performance metric $\mathbf{m} : \mathbb{N}^4 \to \mathbb{R}$. The neighborhood B assigns to each instance a set of nearby instances, based on some context-dependent similarity. It induces two key operations on formulas: *dilation*, which includes instances whose neighborhood intersects the formula, and *erosion*, which retains only those whose neighborhood is fully contained within it. Iterated dilations and erosions allow gradual expansion or contraction of a set of worlds.

Given a training instance φ_ω, the operator $\odot_{(\circ, B, \mathbf{m})}$ maps any classifier $\Psi = (\mathsf{D}, \kappa, \tau)$ to a new classifier $\Psi' = (\mathsf{D}', \kappa', \tau')$ as follows. First, D' is obtained by adding φ_ω to D. Second, κ' is computed by applying $\circ$ to a sequence of formulas induced by φ_ω:

- If $\varphi_\omega \models \mathbf{y}$ (i.e., the instance is positive), the sequence consists of the formula $\varphi_{\omega_{\mathbf{X}}}$ followed by its successive dilations (with respect to B), until a fixed point is reached.
- If $\varphi_\omega \models \neg\mathbf{y}$ (i.e., the instance is negative), the sequence consists of the formula $\neg\varphi_{\omega_{\mathbf{X}}}$ followed by its successive erosions, again until a fixed point is reached.

Last, τ' is chosen to maximize the value of the performance metric $\mathbf{m}$ over D'.

This process locally adjusts the plausibility of the input instance and its neighborhood in line with its label. A positive instance increases the plausibility of nearby instances; a negative one decreases it. The performance metric guides the threshold selection to best separate positive from negative instances.

A concrete learning operator. We instantiated our framework with: the basic shifting operator $\circ_{+1}$, the Hamming-based neighborhood B^H, and the balanced accuracy metric $\mathbf{m}_{ba}$, resulting in the operator $\odot_{(\circ_{+1}, B^H, \mathbf{m}_{ba})}$. We have shown that this operator satisfies a set of interesting properties: it supports polynomial-time training and inference, has a representation-free formulation (classifiers are fully described by the dataset and the threshold), and is robust to instance ordering (permuting training examples yields the same classifier).

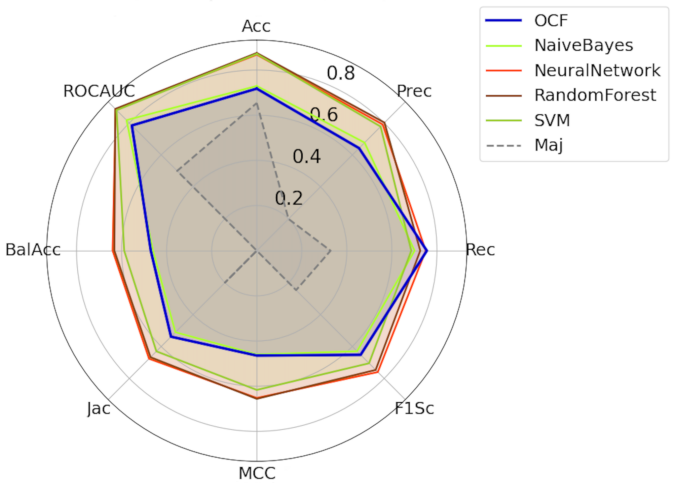

Figure 2: A comparison of performance between our learning model and some standard models across 58 datasets.

To evaluate performance, we compared our framework to nine standard models, including a baseline classifier that always predicts the majority class. The comparison used 58 benchmark datasets from the UCI Machine Learning Repository. Figure 2 presents spider plots comparing six models[1] across eight performance metrics. Our framework (labeled OCF) outperforms the Maj baseline (a minimum requirement for any learning method), matches Naive Bayes, and slightly outperforms other models on recall.

These initial results validate the practical relevance of our framework and highlight the promise of principled learning models grounded in BCT.

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.

Darwiche, A., and Pearl, J. 1997. On the logic of iterated belief revision. *Artificial Intelligence* 89(1-2):1–29.

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

Konieczny, S., and Pino Pérez, R. 2008. Improvement operators. In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR'08)*, 177–187.

Schwind, N.; Inoue, K.; Konieczny, S.; and Marquis, P. 2025. Iterated belief change as learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI'25), to appear*.

[1] Additional results are available in (Schwind et al. 2025).