HAL open science

Less is Better: An Energy-Based Approach to Case Base Competence

Esteban Marquer, Fadi Badra, Marie-Jeanne Lesot, Miguel Couceiro, David Leake

To cite this version:

Esteban Marquer, Fadi Badra, Marie-Jeanne Lesot, Miguel Couceiro, David Leake. Less is Better: An Energy-Based Approach to Case Base Competence. ICCBR ATA'23: Workshop on Analogies: From Theory to Applications – AR & CBR Tools for Metric and Representation Learning at ICCBR2023, Jul 2023, Aberdeen, United Kingdom. pp.27-42. hal-04184905

HAL Id: hal-04184905

https://inria.hal.science/hal-04184905v1

Submitted on 22 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Less is Better: An Energy-Based Approach to Case Base Competence

Esteban Marquer[4,*,†], Fadi Badra[1,*,†], Marie - Jeanne Lesot[3,*,†], Miguel Couceiro[4,*,†] and David Leake[2,*,†]

[1]*Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, LIMICS, Sorbonne Université, INSERM, F-93000, Bobigny, France*

[2]*The Luddy School of Informatics, Computing, and Engineering, Indiana University, USA*

[3]*Sorbonne Université, CNRS, LIP6, F-75005 Paris, France*

[4] *University of Lorraine, CNRS, Loria, Nancy, France*

Abstract

This paper revisits the notion of case base competence in the light of recent advances in the modeling of analogical reasoning, based on the idea of similarity transfer from a situation space to an outcome space. For that we consider the CoAT indicator, that measures the compatibility between two similarity measures on a case base, and use it to define an intrinsic measure of competence of a case base with respect to a reference set. Initial experimental results show that the proposed competence measure correlates with the performance of the CoAT prediction algorithm. In fact, our preliminary results seem to indicate that, under some initial conditions, our competence based model can fit any classification boundary. We then revisit the notions of case competence and locality, and show that some source cases may degrade the overall case base competence while others may improve it, and that a given source case may have disparate influence on different regions of the case space.

Keywords

Competence models, case base compression, energy-based models, case base maintenance, case-based classification

1. Introduction

Case bases are one of the main sources of knowledge used in case-based reasoning (CBR), along with similarity knowledge, adaptation knowledge and domain knowledge [1]. Case acquisition and maintenance therefore constitute crucial steps in the knowledge engineering process of a CBR system. Acquiring and maintaining cases may be expensive, and case storage capacity

ICCBR ATA'23: Workshop on Analogies: From Theory to Applications – AR & CBR Tools for Metric and Representation Learning at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

*Corresponding author.

†These authors contributed equally.

esteban.marquer@loria.fr (E. Marquer); badra@sorbonne-paris-nord.fr (F. Badra); Marie-Jeanne.Lesot@lip6.fr (M. -. J. Lesot); miguel.couceiro@loria.fr (M. Couceiro); leake@indiana.edu (D. Leake)

https://emarquer.github.io/ (E. Marquer); https://limics.fr (F. Badra); https://lip6.fr (M. -. J. Lesot); https://members.loria.fr/mcouceiro/ (M. Couceiro); https://homes.luddy.indiana.edu/leake/ (D. Leake)

0000-0003-2315-7732 (E. Marquer); 0000-0002-2437-8230 (F. Badra); 0000-0002-3604-6647 (M. -. J. Lesot); 0000-0003-2316-7623 (M. Couceiro); 0000-0002-8666-3416 (D. Leake)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3438/paper_03.pdf

may be limited or constrained by selection criteria. Crafting a case base for a given task thus requires addressing questions such as "which cases should be included in the case base?", which can alternatively be expressed as "which cases are the most competent?", where the definition of the competence notion can be seen as the formalization of this issue.

Such questions have been extensively studied in the literature on case base maintenance (*e.g.,* [2, 3, 4, 5, 6, 7, 8]). Most competence models assume that problems are solved by a k-Nearest Neighbor algorithm (often with $k = 1$) augmented by case adaptation, and are strongly influenced by the way this algorithm works. For instance, it is often assumed that only the most similar cases may contribute to solving a new case, provided that they are themselves adaptable. The competence of a case is typically assessed by computing its coverage, *i.e.*, the set of cases that it may contribute to solving. Yet beyond approaches based on the k-Nearest Neighbor algorithm, no case-based prediction algorithm actually complies with this assumption. Algorithms such as CCBI [9], PossIBL [9], or CoAT [10], take into account the similarities with *all* source cases in order to make a prediction. Here we examine competence criteria suitable for such approaches.

Competence methods yield guidance for case deletion to minimize the competence loss in the case base compression process [5, 6]. However, determining a suitable notion of competence over the whole case base is challenging and, to our knowledge, no theoretical guarantees exist to relate the competence of a case base and the performance of a corresponding CBR system.

Recent advances in the modeling of analogical transfer shed new light on the case base competence problem. It has been shown [11] that all case-based prediction methods share the common inference principle based on the transfer of similarity knowledge from the situation space, in which the cases are described, to an outcome space, in which their attached solutions are described. This transfer can then be achieved by optimizing a measure of compatibility between two similarity measures, respectively associated with each of the two spaces. The latter idea motivated the case-based prediction method CoAT [10, 12, 13] that relies on the optimization of a global compatibility indicator between two similarity measures on the case base. In this framework, the predicted outcome is the one that entails the least increase in the value of the CoAT indicator. The latter can be interpreted as an intrinsic indicator of the optimality of the case-based setting for the task at hand. Preliminary results showed it can be used to assess the quality of similarity measures or of solutions.

In this paper, we further explore this indicator to address the problem of case competence. The main idea is to exploit this indicator to define an intrinsic measure of competence of a case base, which could be used later to obtain theoretical guarantees on the link between the competence of a case and the performance of a case-based classifier or predictor. To do so, we first propose to interpret the CoAT global indicator in an energy-based framework [14]. Then the competence of a source case can be intuitively related to its ability to reduce the energy of correct outcomes and to increase the energy of incorrect outcomes. For instance, in a classification setting, this would mean the case is lowering the energy of the good class, and increasing the energy of all others.

This new approach to the problem of case (base) competence has two noteworthy consequences. First, rather than taking the traditional case base maintenance view of considering only the nearest cases, it considers the compatibility of all cases in the case base. Intuitively, this could be important for deletion scenarios, because case bases with lower energy should

provide more stable results when cases are deleted. Second, this makes clear the potential ramification that case deletion could either decrease or increase system performance: cases may be competent (increased overall performance) w.r.t. a given class, while entailing competence degradation (decreased overall performance) w.r.t. another class.

To support the latter and establish the relation between competence and performance of a case-based system, we propose two loss functions (see Sec. 4): one that corresponds to the intuitive notion of competence (*i.e.*, counting positively the energy of correct outcomes and negatively the energy of incorrect ones), and the other inspired by the hinge loss. We perform a comparative study of the two and observe that the latter is preferable to the former. Thus focusing on the hinge loss, we conduct several empirical studies to assess both the performance and robustness of this CoAT-based competence notion (see Sec. 5) in various initial settings. These experiments also indicate the potential use of our competence notion for case base compression and maintenance purposes. Furthermore, they support that our competence-based framework can produce surrogate models capable of approximating different classification boundaries.

The paper is organized as follows. In Sec. 2 we briefly discuss well-known approaches to case base maintenance, and recall key definitions from related work on case competence. The definition of the CoAT indicator is recapped in Sec. 3 and then used in Sec. 4 to propose a new definition of case base competence and of competence of an individual case. We present several empirical studies in Sec. 5 to support our performance and robustness claims, and to analyze the behavior of our approach both quantitatively and qualitatively. Sec. 6 concludes the paper and discusses several perspectives for future work.

Main contributions. The main contributions of the paper are the following:

- We introduce an energy-based framework that relies on the optimization of the CoAT indicator as a measure of similarity compatibility, and which is used to propose new measures of case base competence w.r.t. different loss functions.
- We show empirically that thus defined, the case base competence is tightly linked to the performance of case-based models, which constitutes a promising step towards theoretical guarantees of performance.
- We propose fine-grained competence notions, namely, w.r.t. individual source cases and w.r.t. individual reference cases, which can be used to identify areas of "expertise" of cases in a case-based system.
- We present an empirical study to assess the robustness of the proposed approach w.r.t. to different case base initializations and reference case sets, followed by an iterative and qualitative analysis that shows the potential of the proposed approach for case base maintenance and compression, as well as for fitting any classification boundary.

2. Basic Background and Motivation

This section briefly presents the notation used throughout the paper and recaps the definition of the case base maintenance task in the CBR setting.

2.1. Key Definitions

Let $\mathcal{S}$ denote an input space, and $\mathcal{R}$ an output space. An element of $\mathcal{S}$ is called a *situation*, and an element of $\mathcal{R}$ is called an *outcome*, or a result. A set $CB = \{(s_1, r_1), \ldots, (s_n, r_n)\}$ of elements in $\mathcal{S} \times \mathcal{R}$ is called a *case base*. An element $c = (s, r) \in CB$ is called a *source case*. In addition, the spaces $\mathcal{S}$ and $\mathcal{R}$ are respectively equipped with the similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$, that respectively denote the similarity measure on situations and on outcomes. Let $\mathcal{T} \subset \mathcal{S} \times \mathcal{R}$ be a set of cases called a *reference set*, and $c_t = (s_t, r_t) \in \mathcal{T}$ be a reference case. We will write $(s_t, \hat{r})$ to denote a potential case constructed by keeping the same situation $s_t \in \mathcal{S}$, but choosing a different outcome $\hat{r} \in \mathcal{R}, \hat{r} \neq r_t$ for the case.

2.2. Case Base Maintenance

Case base maintenance revises the contents or organization of a case base to improve performance, and is a longstanding research area for CBR (*e.g.*, [15]). Much of this work has studied case base compression by case deletion [5]. Compression efforts were initially motivated by the desire to control retrieval costs and respect storage constraints. Advances in computational power have reduced some of these concerns in practice [16] but compression remains useful when efficiency is paramount and for reasons such as reducing the number of cases for a knowledge engineer to maintain. Deletion of cases may remove the knowledge required to solve particular problems, motivating maintenance work focused on retention of case base competence, which Smyth [6] and McKenna define as the range of problems a CBR system can successfully solve.

Case base compression strategies are often deletion-based, aimed at successively removing cases whose loss will least harm competence. Estimates of case competence contributions are commonly done based on the existing cases in the case base, under the representativeness assumption [6] that the case base is a good predictor of the distribution of future problems. This assumption is expected to hold for domains well suited to CBR, when the case base is sufficiently mature, though may be endangered by problem drift (*e.g.*, [17]). Estimation of expected case competence contributions is commonly based on considering relationships between cases and their nearest neighbors in the case base, favoring cases that have high coverage of other cases and low reachability, *i.e.*, that are recoverable from fewer cases [6].

This paper presents a maintenance perspective that is novel in three ways. First, rather than emphasizing the relationship of cases to nearby neighbors, the core of the approach is a global optimization of a case base energy function. Second, rather than using a global approximation of future problems, it defines competence with respect to specific reference sets. Third, it questions the assumption that case base compression entails competence loss and illustrates that compression may actually enhance performance, providing a new motivation for case base compression.

3. The CoAT Method

The CoAT method [10, 12, 13] performs analogical transfer by minimizing a global indicator of compatibility between two similarity measures. In this section, we recall the definition of this

indicator, and show that it can be seen as an energy function.

3.1. Definition of the CoAT Indicator

The compatibility of $\sigma_{\mathcal{R}}$ with $\sigma_{\mathcal{S}}$ for a given case base CB is measured globally on the case base CB, by a global indicator denoted $\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$. The latter takes an ordinal point of view on the whole case base CB, by checking if the order induced by $\sigma_{\mathcal{R}}$ is the same as the one induced by $\sigma_{\mathcal{S}}$. The following continuity constraint is tested on each triple of cases (c_0, c_i, c_j), with $c_0 = (s_0, r_0)$, $c_i = (s_i, r_i)$, and $c_j = (s_j, r_j)$:

$$\text{if } \sigma_{\mathcal{S}}(s_0, s_i) \geq \sigma_{\mathcal{S}}(s_0, s_j), \text{ then } \sigma_{\mathcal{R}}(r_0, r_i) \geq \sigma_{\mathcal{R}}(r_0, r_j). \tag{C}$$

Constraint (C) expresses that whenever a situation s_i is more similar to situation s_0 than to situation s_j, this order should be preserved on outcomes. A triple (c_0, c_i, c_j) does *not* satisfy (C) if case c_i is more similar to case c_0 than to case c_j for situations, but less similar for outcomes, *i.e.*, when $\sigma_{\mathcal{S}}(s_0, s_i) \geq \sigma_{\mathcal{S}}(s_0, s_j)$ and $\sigma_{\mathcal{R}}(r_0, r_i) < \sigma_{\mathcal{R}}(r_0, r_j)$. Such a violation of the constraint is called an *inversion of similarity*. The indicator $\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$ counts the total number of inversions of similarity observed on a case base CB:

$$\begin{aligned}\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB) = |\{((s_0, r_0), (s_i, r_i), (s_j, r_j)) \in CB \times CB \times CB \text{ such that} \\ \sigma_{\mathcal{S}}(s_0, s_i) \geq \sigma_{\mathcal{S}}(s_0, s_j) \text{ and } \sigma_{\mathcal{R}}(r_0, r_i) < \sigma_{\mathcal{R}}(r_0, r_j)\}|.\end{aligned}$$

For a new situation s_t, the transfer inference consists in finding the outcome r_t that leads to the new case $c_t = (s_t, r_t)$ that minimizes the value of the indicator:

$$r_t = \arg\min_{r \in \mathcal{R}} \Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB \cup \{(s_t, r)\}). \tag{1}$$

An important aspect to notice is that the CoAT method makes use of the whole case base CB, and not only the most similar case(s), in order to predict the outcome of the new case.

3.2. An Energy Function View on the CoAT Method

After briefly reminding the principles of the energy-based framework for solving machine learning tasks, this section proposes to interpret the CoAT optimization of the global indicator in this setting.

Energy-based models. Inspired from statistical physics, energy-based models [14] specify a probability distribution $p(x; \theta) = e^{-E_\theta(x)} / \int e^{-E_\theta(x)} dx$ via a parameterized scalar-valued function $E_\theta(x)$ called an *energy function*. In its conditional version, the definition of an energy function $E_\theta : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ associates to each pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ a scalar value $E_\theta(x, y)$ that represents the compatibility between the input x and the output y under the set of parameters θ. The energy function E_θ takes low values when y is compatible with x, and higher values when y and x are less compatible. The goal of the energy-based *inference* is to find, among a set of outputs $\mathcal{Y}$, the output $y^* \in \mathcal{Y}$ that minimizes the value of the energy function:

$$y^* = \arg\min_{y \in \mathcal{Y}} E_\theta(x, y).$$

Given a family of energy functions $E_\theta(x, y)$ indexed by a set of parameters θ, the goal of the *learning* step is to optimize the θ parameters in order to "push down" (*i.e.*, assign lower energy values to) the points on the energy surface that are around the training samples, and to "pull up" all other points. Contrastive divergence [18] is a common learning strategy that, given a numerical hyperparameter λ, consists in optimizing a contrastive loss function such as the hinge loss, which is defined, for a training sample (x_k, y_k) and a generated out of distribution sample $(x_k, \hat{y})$ by: $\ell(\theta, x_k, y_k, \hat{y}) = \max(0, \lambda + E_\theta(x_k, y_k) - E_\theta(x_k, \hat{y}))$. The hinge loss associates a loss value to a training sample (x_k, y_k) whenever its energy is not lower by at least a margin λ than the energy of the incorrect sample $(x_k, \hat{y})$.

The CoAT indicator as an energy function. The CoAT case-based prediction method can be interpreted in the energy-based model framework, in which the energy $E_\theta(s_t, r)$ of any new case (s_t, r) is given by the value of the Γ indicator when the case is added to the case base, *i.e.*,

$$E_\theta(s_t, r) = \Gamma(\sigma_\mathcal{S}, \sigma_\mathcal{R}, CB \cup \{(s_t, r)\}).$$

The input space $\mathcal{X}$ is the situation space $\mathcal{S}$ and the output space $\mathcal{Y}$ is the outcome space $\mathcal{R}$. The energy function $E_\theta : \mathcal{S} \times \mathcal{R} \longrightarrow \mathbb{R}$ measures the compatibility of the outcome similarities with the added situation similarities when the potential new case $\hat{c}_t = (s_t, r)$ is added to the case base. The energy function E_θ is parameterized by $\theta = (\sigma_\mathcal{S}, \sigma_\mathcal{R}, CB)$ which includes the case base CB. The goal of the energy-based *inference* is to find, among the set of potential outcomes $r \in \mathcal{R}$, the outcome r_t that minimizes the value of the energy function, and (1) can be reformulated as:

$$r_t = \arg\min_{r \in \mathcal{R}} E_\theta(s_t, r).$$

4. Measuring Competence

In this section we introduce new case (base) competence measures using the previous energy-based framework of the CoAT indicator, w.r.t. different loss functions. We then propose fine-grained variants of competence, at individual and reference case levels.

4.1. Idea of the Method

In the CoAT energy-based model, the energy function $E_\theta(s_t, r)$ is used to compute a (scalar) energy value for each potential outcome r of the new case c_t. The difference between the energy of the predicted outcome and the lowest energy of all other outcomes can be interpreted as a measure of prediction confidence. Therefore, our goal is to capture the idea that the competence of a case base should be related to its ability to maximize the prediction confidence, by decreasing the energy of the correct outcome of a new case and increasing the energy of incorrect outcomes.

We consider two different loss functions of the underlying energy-based model that take as input, besides the $\theta = (\sigma_\mathcal{S}, \sigma_\mathcal{R}, CB)$ parameters of the energy function, that are considered to be fixed, an auxiliary set of reference cases $\mathcal{T}$. The intuition is that, if $\sigma_\mathcal{S}$ and $\sigma_\mathcal{R}$ are fixed, optimizing such a loss function should allow us to learn the right case base CB for the task, *i.e.*, address the case base maintenance issue.

4.2. Competence of a Case Base

This section discusses two definitions of the competence of a case base CB with respect to a reference set $\mathcal{T}$, that are defined from two different loss functions of the energy-based model.

MCE loss competence. The first definition of competence we propose, denoted C_{MCE} relies on the notion of the minimum classification error loss ℓ_{MCE} [14] classically used in the energy-based framework. More precisely, C_{MCE} computes the average value, across the reference set, of this loss ℓ_{MCE} that is defined as the difference between the energy of the correct outcome and the minimum energy of a reference case if it were assigned a different outcome:

$$C_{MCE}(CB, \mathcal{T}) \quad = \quad -\frac{1}{|\mathcal{T}|} \sum_{c_t \in \mathcal{T}} \ell_{MCE}(CB, c_t),$$

where $\ell_{MCE}(CB, c_t) = E_\theta(s_t, r_t) - \min_{\hat{r} \neq r_t} E_\theta(s_t, \hat{r})$. For a correctly classified instance, $\ell_{MCE}(CB, c_t)$ is a negative value whose magnitude can be interpreted as the prediction confidence of CoAT, as mentioned previously. For an incorrectly classified instance, $\ell_{MCE}(CB, c_t)$ is a positive value that allows to measure the extent of the error, *i.e.*, how much the true class is missed. As a consequence, the lower the $\ell_{MCE}(CB, c_t)$ value, the better and, due to the - sign in the definition of $C_{MCE}(CB, \mathcal{T})$, the greater the $C_{MCE}(CB, \mathcal{T})$, the better, *i.e.*, the more competent CB is w.r.t. $\mathcal{T}$.

Hinge loss competence. The hinge loss competence C_{hinge} modifies the minimum classification error loss by integrating an additional parameter, denoted by λ, that corresponds to a margin. The values of $\ell_{MCE}(CB, c_t)$ that are lower than $-\lambda$ (corresponding to the instances with high prediction confidence) are not taken into account and not allowed to compensate for the misclassified instances:

$$C_{hinge}(CB, \mathcal{T}) \quad = \quad -\frac{1}{|\mathcal{T}|} \sum_{c_t \in \mathcal{T}} \ell_{hinge}(CB, c_t),$$

where $\ell_{hinge}(CB, c_t) = \max(0, \lambda + \ell_{MCE}(CB, c_t))$.

Comparison between the competence metrics. C_{MCE} is close to a direct translation of the notion of competence described in Subsec. 4.1. However, in C_{MCE}, the negative contributions to competence (incorrect predictions) and the positive ones (correct predictions) can cancel each other out. In other words, a high increase of the confidence for correctly predicted class can compensate for a lot of small misclassifications. This scaling issue between negative and positive contributions to C_{MCE} is avoided in C_{hinge} as only the negative contributions (to a margin) are accounted for.

4.3. Fine-Grained Competence: Case Level and Expertise Areas

This section proposes to break down the case base competence at a more refined level, considering the individual source and reference cases levels.

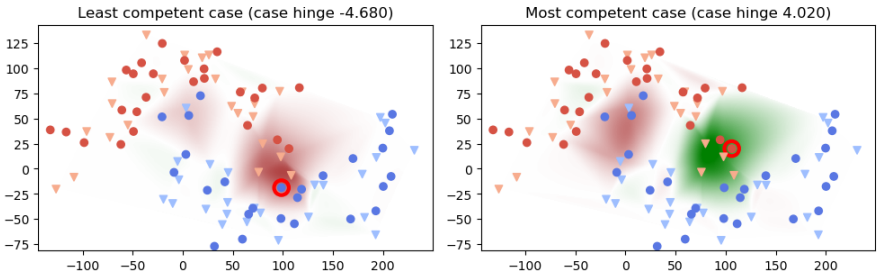


Figure 1: Influence map of 2 source cases c_1 and c_2 (circled in red) of the Half Moon dataset (CB=colored disks, $\mathcal{T}$=pale colored triangles): the background color shows, at each position x, y, the value of $influence(c_1, (x, y))$, where green corresponds to a positive value and red to a negative one.

Proposed definitions. We first propose to define the competence of a source case $c = (s, r) \in CB$ w.r.t. a reference set $\mathcal{T}$ as the loss of competence that would happen if this source case was deleted from the case base:

$$C(c, CB, \mathcal{T}) = C(CB, \mathcal{T}) - C(CB \setminus \{c\}, \mathcal{T}).$$

As for any competence measure, the greater, the better.

At an even finer level, we define the notion of competence locally as the contribution of a source case $c = (s, r) \in CB$ on each individual reference case $c_t \in \mathcal{T}$. Indeed, the competence $C(c, CB, \mathcal{T})$ over the reference set $\mathcal{T}$ equals the sum over $\mathcal{T}$ of the loss (*e.g.*, ℓ_{hinge} or ℓ_{MCE}): the above-defined competence of a case $c \in CB$ w.r.t. $\mathcal{T}$ can be expressed as

$$C(c, CB, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{c_t \in \mathcal{T}} influence_{CB}(c, c_t)$$

where $influence_{CB}(c, c_t) = \ell(CB, c_t) - \ell(CB \setminus \{c\}, c_t)$. This notion of case influence entails the idea of locality: all source cases contribute to the competence of the case base, but each source case may contribute differently on different regions of space. This enables the identification of *regions of expertise* of a source case. Since the loss distinguishes between correct and incorrect classification, case influence can also be used to identify those regions where a source case can improve the performance from those where performance is degraded.

Illustrative example. Figure 1 offers a visualization of the case competence and influence considering two source cases, circled in red, of the Half Moon dataset (see Subsec. 5.1), and the map of their influence values. The left figure shows the least competent source case ($C(c, CB, \mathcal{T}) = -4.680$) which degrades the performance on many reference cases (dark red regions) and contributes negatively to the overall competence. This case would be the first one to be removed in a case deletion strategy. The right figure shows the most competent case ($C(c, CB, \mathcal{T}) = 4.020$), which contributes positively to the competence of the case base: it improves the performance of a large set of reference cases (green regions). Interestingly, this case also harms the performance for some references of the opposing class.

Algorithm 1 Case deletion procedure

Require: An initial case base CB and a reference set $\mathcal{T}$
  while $|CB| > 0$ **do**
    $c_{worse} = \arg\min_{c \in CB} C(c, CB, \mathcal{T})$
    $CB = CB \setminus \{c_{worse}\}$
  end while

Case deletion procedure. Source case competence can be applied in a case deletion procedure, as described by Algorithm 1: at each iteration, the source case c_{worse} that contributes least to the competence of the case base CB w.r.t. the reference set $\mathcal{T}$ is deleted from the case base. To observe the effects of successive deletions this algorithm is exhaustive, but in practice deletion would repeat only until a stopping criterion is reached (*e.g.*, desired compression).

5. Experiments

We investigate experimentally the properties of the proposed case deletion procedure and competence definitions, in particular examining their correlation with the classification performance of the CoAT prediction algorithm. We also provide a stability analysis, as well as a qualitative analysis of the results.

5.1. Considered Artificial Datasets

The experiments are performed in a binary classification setting with three synthetic 2D datasets generated from 3 distributions coined Line, Ring, and Half Moon and respectively illustrated in Fig. 2a, 2b, and 2c.

The Line data are drawn from a uniform distribution defined on $[0, 2] \times [0, 3]$. They are labeled according to the arbitrary chosen line $f(x) = -x + 2.5$, and noise is added by randomly switching the label, with a probability of 20%, for cases within a 0.3 distance to the boundary.

For the Ring data, two classes are defined a concentric rings of radii 25 and 50. For each class, points are randomly sampled using polar coordinates, drawing the angle from a uniform distribution on $[0, 2\pi]$ and radius from a normal distribution $\mathcal{N}(\mu = r_c, \sigma = 10)$, where $r_c \in \{25, 50\}$ is the radius of the class. The theoretical decision boundary for the Ring data is the circle of radius 32.5.

The Half Moon dataset is generated with "make_moons" function from the Scikit-Learn library[1] with a noise of 0.2. The distribution is composed of two halves of a circle, one of which is shifted laterally by the radius. Each half-circle corresponds to a class.

In all three cases, $\mathcal{S} = \mathbb{R}^2$ and the associated similarity is a decreasing function of the standard Euclidean distance $\sigma_{\mathcal{S}}(x, y) = \exp(-d^2(x, y))$; the outcome space is $\mathcal{R} = \{0, 1\}$ equipped with $\sigma_{\mathcal{R}}(r_x, r_y) = 1$ if $r_x = r_y$ and 0, otherwise. Note that the three data distributions are more or less compatible with $\sigma_{\mathcal{S}}$ due to their geometry. In that regard, the limitations of $\sigma_{\mathcal{S}}$ help understand the performance of our approach when the similarity is not as good as it could be.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html

5.2. Experimental Protocol

For each of the three data distributions considered, we generate 1000 samples that we split into 20 non-overlapping subsets of 50 cases, each one being balanced in terms of classes. We separate them in 2 groups: 10 serve as initial case bases $CB_1, \ldots, CB_{10}$ and the others as reference sets $\mathcal{T}_1, \ldots, \mathcal{T}_{10}$. Fig. 2a, 2b, and 2c display the overall 1000 samples together with their label, showing in light colors the reference sets.

For each pair $(CB_i, \mathcal{T}_j)$, we apply the proposed compression algorithm. After each removal step, the classification results obtained by the CoAT algorithm applied with the current CB are assessed by the macro F1 on all reference cases $\bigcup_{k=1..10} \mathcal{T}_k$.

Fig. 2d, 2e and 2f show the evolution of this macro F1 criterion during the case deletion procedure, comparing the two proposed competence measures $C_{MCE}(c, CB, \mathcal{T})$ and $C_{hinge}(c, CB, \mathcal{T})$; the shade corresponding to the 95% confidence interval over the 100 combinations of initial case base and references. In Fig. 2g, 2h, and 2i, each line shows the results for one of the 10 case bases and the shades show the 95% confidence interval over the 10 reference sets. Reciprocally, in Fig. 2j, 2k, and 2l, each line corresponds to a reference set and the shades correspond to the 95% confidence interval over the 10 initial case bases.

5.3. Results

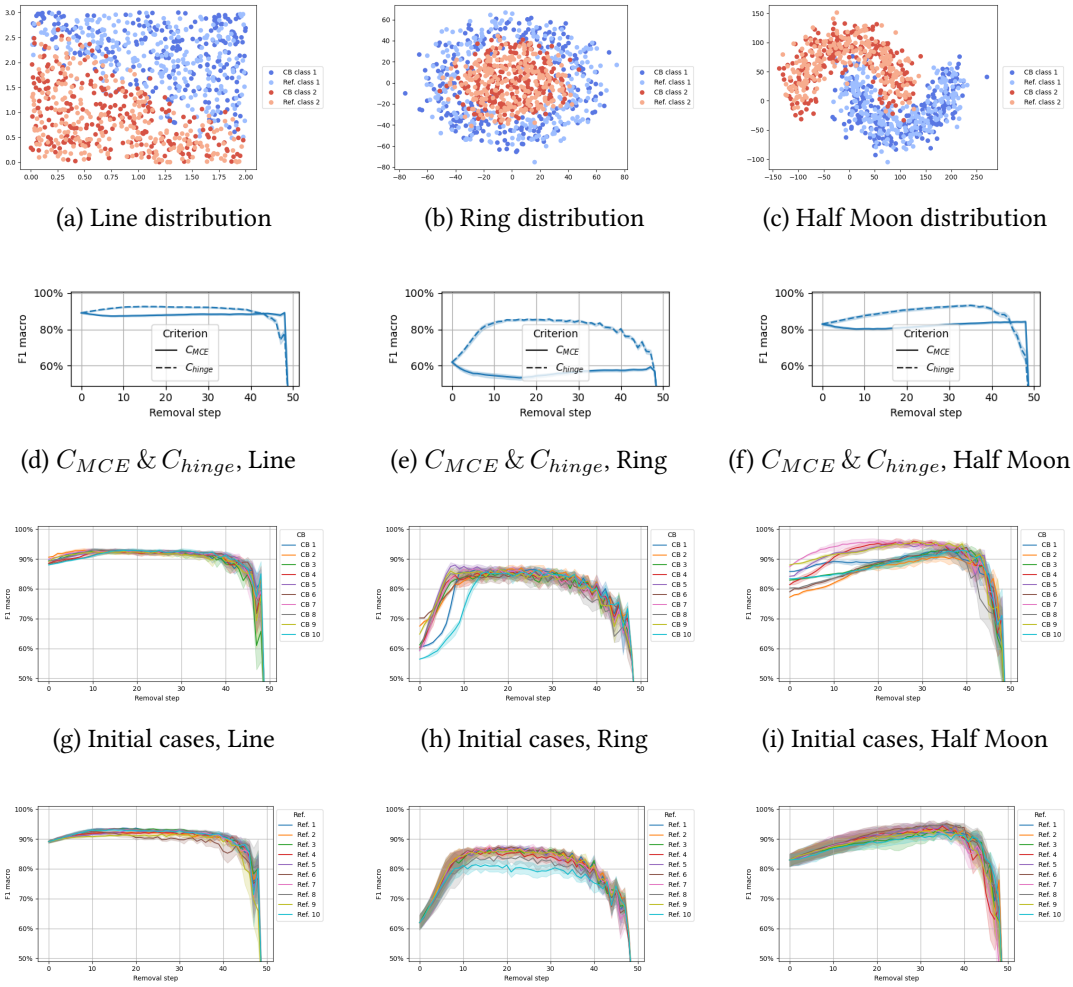
We study the behavior of the case-based models resulting from compression by performing both quantitative and qualitative analyses.

5.3.1. Competence Definitions and Correlation with Performance

In the second row of Fig. 2, we compare the evolution of the macro F1 when using either C_{MCE} or C_{hinge} for case competence in the compression process. With C_{MCE}, F1 remains at its maximum slightly longer, so a few more cases can be removed. However, with C_{MCE}, F1 remains at its initial value throughout the process and does not reach as high values as with C_{hinge}. For instance, on Ring, F1 reaches a value close to 60% for C_{MCE} and 85% for C_{hinge}.

Complementary experiments, whose curves are omitted for brevity, examine the evolution of the case base competence during the compression process. They show that, as desired, the case base competence remains constant or increases during compression when using C_{hinge}. On the other hand, using C_{MCE} causes an increasingly faster decrease of competence, making it a poor choice for case base compression. Also, by comparing the decrease when using C_{MCE} with F1, which is almost constant, it becomes striking that C_{MCE} is not directly correlated with predictive performance, and this is problematic in our vision of competence. As mentioned in Subsec. 4.2, prediction successes and failures are considered at the same time in C_{MCE}, but higher C_{MCE} could be an expression of higher confidence in already well predicted cases, of fewer errors, or of less confident errors. In that regard, C_{hinge} is more suitable as a competence measure as it measures how confident the model is in its errors, and thus higher C_{hinge} corresponds to fewer or less confident errors, which directly translates to higher performance.

The experiments described hereafter consider only C_{hinge}, as it is more interesting in terms of performance, stability across datasets, and is a better fit for the notion of competence.



(a) Line distribution (b) Ring distribution (c) Half Moon distribution

(d) C_{MCE} & C_{hinge}, Line (e) C_{MCE} & C_{hinge}, Ring (f) C_{MCE} & C_{hinge}, Half Moon

(g) Initial cases, Line (h) Initial cases, Ring (i) Initial cases, Half Moon

(j) Reference cases, Line (k) Reference cases, Ring (l) Reference cases, Half Moon

Figure 2: Evolution of the macro F1 (on all 500 reference cases) during the case deletion procedure, for Line, Ring and Half Moon. The distribution of the 1000 cases used is displayed in first row. In the second row, performance with C_{MCE} is compared with C_{hinge} (C_{MCE} & C_{hinge}). The performance with C_{hinge} is also detailed when grouping by case base initialization (third row) and reference set (fourth row).

5.3.2. Competence of the Case and Impact on Performance

Looking at the F1 evolution (see Fig. 2g to 2l) shows that no matter the initial cases in the case base or the reference cases used, the same trend of performance can be observed: *(i)* a raise, *(ii)* a plateau, and finally *(iii)* a faster and faster decrease. As our algorithm removes the cases by order of increasing competence, it appears that *(i)* corresponds to incompetent cases, *(ii)* to cases that are neither competent nor incompetent, and *(iii)* to competent cases. Going further, during phase *(i)* removing cases improves the performance, meaning that the removed cases were "polluting" the case base. In *(ii)*, the removed cases neither harm nor benefit the performance of

the case base, as such they can be considered redundant w.r.t. the remaining cases. The cases that remain are the most competent and useful ones, and in *(iii)* they are removed by order of increasing competence, leading to sharper and sharper drops in performance.

This behavior is similar to the footprint deletion procedure from Smyth and Keane [5], with the auxiliary, spanning, and support cases removed in *(ii)*, and pivotal cases removed in *(iii)*. Compared to [5], our procedure is more powerful as it can handle case bases that do not properly fit the distribution of the data, as harmful cases are removed in priority in *(i)*.

As we observe a striking parallel between the performance change and the compression step, it appears that C_{hinge} suits the intuition of competence, since the step at which a case is removed during compression is proportional to its competence. Furthermore, this general trend provides empirical guarantees that the maximum performance is reached just before the first significant decrease in performance, meaning we can stop the process as soon as we detect such a decrease.

5.3.3. Robustness of the Compression

Robustness w.r.t. the initialization of the case base. Fig. 2g, 2h, and 2i show that the initial cases in the case base change the initial performance and time needed to converge to the general trend of performance. In extreme cases of poor initial performance, the convergence might be delayed until after performance starts to decrease, as can be seen in Fig. 2i for the lower of the two groups of case bases.

By analyzing the distribution of each set of initial cases (not shown here for brevity), we observe that not having enough cases in a particular area of the distribution (*i.e.*, having holes in the case base in important places) causes the case base to have difficulties to reach the best performance. We were able to confirm this effect by manually removing cases in parts of the distribution, in experiments omitted here for brevity. Conversely, if we manually make one class over-represented, the performance is not damaged as much, as the cases in the over-represented class are redundant and are removed in the plateau *(ii)*.

From these results, the initial cases harm the best performance only when the initial performance is too poor (leading to converging too slowly to reach the best state) or when there are no cases in an important area of the boundary.

Robustness w.r.t. the reference cases. The cases used to measure the competence can have a critical impact on the best performance reached. If there is no major gap between the distribution of references and the true distribution of the data, the maximal performance can be harmed but is still in the same range as the other references, as can be seen with the cyan references in Fig. 2g and 2h. However, the effect of the references becomes striking when we manually create holes in the distribution of references, in experiments omitted for brevity. In that setting, the case base becomes biased towards the incorrect distribution of the references.

5.3.4. Qualitative Analysis

Fig. 3 displays, for each dataset and for a single initialization and reference, 3 steps of the case deletion procedure: the initial case base (first column), after 10 deletion steps (second column),

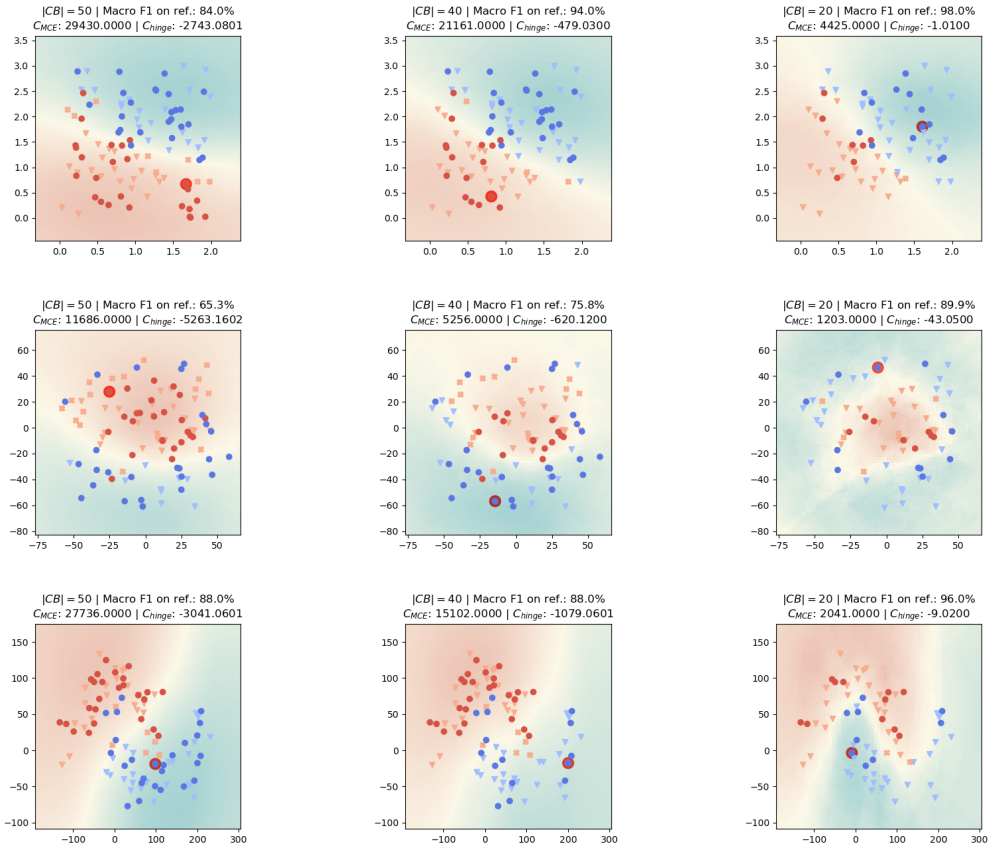


Figure 3: Three steps (first column: initial step, middle column: after 10 deletions, third column: after 30 deletions) of the case deletion procedure, for different datasets (top row: Line, middle row: Ring, bottom row: Half Moon). The case circled in red is the one that will be removed from the case base.

and after 30 deletion steps (third column). In each figure, the red and blue dots represent the remaining source cases, and the crosses and the triangles represent the references (triangles and crosses respectively mean correct and incorrect prediction). The least competent source case c_{worse} that will be deleted is circled in red.

The colored map in the figure represents CoAT's predictions for new cases across the space, with the color matching the predicted class and the saturation corresponding to the confidence (*i.e.*, the energy difference between the two outcomes, see Subsec. 4.1). In that manner, it is possible to identify the decision boundary of the compressed case base. At the end of the process, CoAT's decision frontier meets the theoretical classification boundary of the distribution, even for Half Moon, which has a relatively complex boundary. The decision frontier induced by the compressed case base is thereby able to closely approximate the ideal classification boundary.

5.4. Discussion

The compression process using C_{hinge} is able to reduce the number of cases in the case base to 40% (Ring) or even 20% (Line and Half Moon) of its initial size, while strictly improving performance. While our current experiments only cover binary classification, our approach is designed to handle any kind of nominal data in the outcome space. Further work on the approach will include multi-class classification and real-world data.

The robustness experiments show that the initial case base is not a major factor in the peak performance, as long as there are enough cases in the important regions of the situation space. However, it is important to have a proper set of reference cases, as the distribution of the references is closely matched by the compressed case base. If the reference cases are not representative of the true distribution of the data, then the compressed case base is not guaranteed to match the true distribution. To summarize, it is useful to focus on the quality of the reference (*i.e.*, how representative of the actual distribution they are) and on having sufficient initial cases for the case base, even if their quality is rather poor, as long as they cover enough of the distribution for the intended purpose of the model.

Additionally, we obtain empirical evidence of the benefits of C_{hinge} over C_{MCE}, and the performance of the case base measured by C_{hinge} correlates to CoAT's prediction performance. The question of whether this measure of competence is compatible with other CBR processes than CoAT remains open, in particular since CoAT and our competence measure are based on the same energy function. The ordering of cases—based on their competence—may change after a case is removed, as our competence measure involves the rest of the case base. This might have an effect on the compression process, but our energy-based approach to competence may offer theoretical guaranties or bounds on those changes. If the competence of a case remains stable when removing another case, we can speed up convergence by removing cases by batches.

6. Conclusion and Future Work

This paper introduced an energy-based approach to measuring the competence of a case base for machine learning tasks such as case prediction and classification. This competence approach differs from prior approaches proposed in the literature as it relies on the optimization of a global compatibility indicator between two similarity measures, one on the situation space and the other on the outcome space.

We show empirically that this notion of competence is tightly related to performance for a case-based classification task, in the sense that the competence of a source case is positively correlated to its ability to reduce the energy of correct outcomes and to increase the energy of incorrect outcomes. We analyze both quantitatively and qualitatively the behavior of this competence-based approach on different datasets (with substantially different distributions) and taking into account different classification frontiers and loss functions. Moreover, we analyze its robustness with respect to different reference and initial cases.

These results suggest the strong potential of this energy-based framework for guiding case base maintenance, providing an alternative to existing methods. One of the main differences is that it employs a global approach by considering the competence of a case base as a whole, rather than a local approach as it is often the case in the literature (where only nearest neighbors

are considered). The empirical and thorough comparison between the former and the latter will constitute one of the topics to be investigated in a future contribution.

7. Acknowledgments

This research was partially supported by the ANR project "Analogies: from theory to tools and applications" (AT2TA), ANR-22-CE23-0023 and by the ANR project "Similarity Measure Learning for Analogical Transfer" (SMeLT), ANR-22-CE23-0032. David Leake's work was funded in part by the Department of the Navy, Office of Naval Research (Award N00014-19-1-2655).

References

[1] M. M. Richter, Knowledge Containers, in: Readings in Case-Based Reasoning, 2003. URL: https://www.researchgate.net/publication/225070310_Knowledge_Containers.

[2] N. Arshadi, I. Jurisica, Maintaining case-based reasoning systems: A machine learning approach, in: P. Funk, P. A. González-Calero (Eds.), Advances in Case-Based Reasoning, 7th European Conference, ECCBR'04', volume 3155 of *Lecture Notes in Computer Science*, Springer, 2004, pp. 17–31.

[3] L. Cummins, D. Bridge, On Dataset Complexity for Case Base Maintenance, in: A. Ram, N. Wiratunga (Eds.), Case-Based Reasoning Research and Development, volume 6880, Springer, 2011, pp. 47–61.

[4] L. Cummins, Combining and Choosing Case Base Maintenance Algorithms, Ph.D. thesis, University College Cork, 2013.

[5] B. Smyth, M. Keane, Remembering to forget: A competence-preserving case deletion policy for case-based reasoning systems, in: Proc. of the 13th Int. Joint Conf. on Artificial Intelligence IJCAI, Morgan Kaufmann, 1995, pp. 377–382.

[6] B. Smyth, E. McKenna, Competence models and the maintenance problem, Computational Intelligence 17 (2001) 235–249.

[7] S. C. K. Shiu, D. S. Yeung, C. H. Sun, X. Wang, Transferring case knowledge to adaptation knowledge: An approach for case-base maintenance, Comput. Intell. 17 (2001) 295–314.

[8] J. Zhu, Q. Yang, Remembering to add: Competence-preserving case-addition policies for case base maintenance, in: Proc. of the 15th Int. Joint Conf. on Artificial Intelligence, IJCAI, Morgan Kaufmann, 1999, pp. 234–241.

[9] E. Hüllermeier, Credible case-based inference using similarity profiles, IEEE Transactions on Knowledge and Data Engineering 19 (2007) 847–858.

[10] F. Badra, A Dataset Complexity Measure for Analogical Transfer, in: Proc. of the 29th Int. Joint Conf. on Artificial Intelligence IJCAI, 2020, pp. 1601–1607.

[11] F. Badra, M.-J. Lesot, Case-Based Prediction – A Survey, IJAR 158 (2023) 108920.

[12] F. Badra, M.-J. Lesot, Theoretical and Experimental Study of a Complexity Measure for Analogical Transfer, in: ICCBR, 2022, pp. 175–189.

[13] F. Badra, M.-J. Lesot, CoAT-APC: When Analogical Proportion-based Classification Meets Case-Based Prediction, in: Proc. of the workshop on Analogies: from Theory to Applications ATA@ICCBR, CEUR-WS, 2022.

[14] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. J. Huang, A Tutorial on Energy-Based Learning, in: Predicting Structured Data, MIT Press, 2006.
[15] D. Wilson, D. Leake, Maintaining case-based reasoners: Dimensions and directions, Computational Intelligence 17 (2001) 196–213.
[16] G. Houeland, A. Aamodt, The utility problem for lazy learners - towards a non-eager approach, in: I. Bichindaritz, S. Montani (Eds.), Case-Based Reasoning Research and Development, ICCBR 2010, Springer, 2010, pp. 141–155.
[17] D. Leake, B. Schack, The problem drift problem and first steps towards addressing it for CBR, in: Case-Based Reasoning Research and Development, ICCBR 2023, Springer, 2023. In press.
[18] G. Hinton, S. Osindero, M. Welling, Y.-W. Teh, Unsupervised Discovery of Nonlinear Structure Using Contrastive Backpropagation, Cognitive Science 30 (2006) 725–731.