

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

Esteban Marquer¹, Fadi Badra², Marie-Jeanne Lesot³,
Miguel Couceiro¹, David Leake⁴

¹ Université de Lorraine, CNRS, Loria, Nancy, France

² Université Sorbonne Paris Nord, LIMICS, Sorbonne Université, INSERM,
F-93000, Bobigny, France

³ Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

⁴ The Luddy School of Informatics, Computing, and Engineering, Indiana
University, USA

July 17th 2023

Outline

- 1 Preliminaries
 - Case Based Reasoning
 - CoAT indicator
 - Energy-based models

- 2 Contributions
 - CoAT and energy-based models
 - Competence using CoAT
 - CB maintenance experiments

Case-Based Reasoning and Analogical Transfer

Case-Based Reasoning (CBR):

- Situations S
- Outcomes R
- Case (s_i, r_i) : situation and corresponding outcome
- Known cases: Case Base (CB)
- Task: given s_j , find its outcome

For 1-nearest neighbor, we take $(s_i; r_i) \in CB$ such that:

	s_j (new)	most similar to s_i
then	r_j (to predict)	most similar to r_i

$$s_i : s_j :: r_i : r_j$$

CoAT: Complexity-based Analogical Transfer

CoAT (Badra 2020)¹: Complexity-based Analogical Transfer

(Dis)similar situations have (dis)similar outcomes

Transfer similarity: situations outcomes

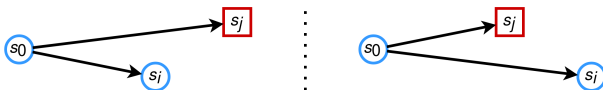
We need compatible similarity measures

¹Fadi Badra (2020). "A Dataset Complexity Measure for Analogical Transfer". In: *Proc. of the 29th Int. Joint Conf. on Artificial Intelligence IJCAI*, pp. 1601–1607

CoAT: Complexity-based Analogical Transfer

If s_0 closer to s_i than to s_j ,
then r_0 closer to r_i than to r_j .

σ_S : Euclidean similarity



σ_R : class membership

r_0 r_i close

r_0 r_j far

σ_S and σ_R agree

r_0 r_i close

r_0 r_j far

σ_S and σ_R disagree

CoAT: Complexity-based Analogical Transfer

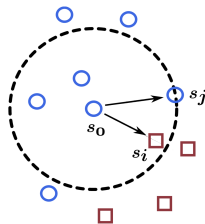
Continuity constraint: Ordinal compatibility between σ_S, σ_R

if $\sigma_S(s_0, s_i) < \sigma_S(s_0, s_j)$ then $\sigma_R(r_0, r_i) < \sigma_R(r_0, r_j)$

Inversion of similarity:

$\sigma_S(s_0, s_i) < \sigma_S(s_0, s_j)$ but $\sigma_R(r_0, r_i) > \sigma_R(r_0, r_j)$

CoAT indicator (σ_S, σ_R, CB) : total number of inversions of similarity observed on a case base CB



CoAT: Complexity-based Analogical Transfer

Compatibility of σ_S, σ_R with the CB: no inversion = perfect match

Predicting:

$$r_t = \arg \min_{r \in R} (\sigma_S, \sigma_R, CB [f(s_t, r)g])$$

Energy-based models: macroscopic view on probability

“An **energy-based model** is a **probabilistic model** governed by an energy function that describes the probability of a certain state. [...] The Boltzmann distribution [(thermodynamics)] establishes a concrete relationship between energy and probability: **low-energy states are the most likely to be observed.**”¹

Energy function E

Learning model: low energy for correct outcome, high for other

Predicting: given a situation, find outcome with the lowest energy

$$r_t = \arg \min_{r \in R} E(s_t, r)$$

¹The Physics of Energy-Based Models. Patrick Huembeli, Juan Miguel Arrazola, Nathan Killoran, Masoud Mohseni, Peter Wittek, Towards data science, 2021.

CoAT in terms of energy

Energy function & : global view
 low energy/ : more desirable outcome

Energy of a new case:

$$E(s_t, r) = (\sigma_S, \sigma_R, CB [f(s_t, r)g)$$

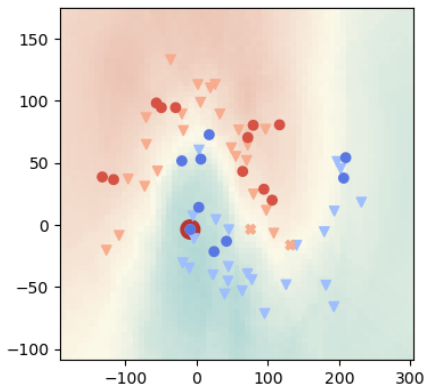
(s_t, r) : new case added to the CB

$E : S \times R \rightarrow \mathbb{R}$, with $\theta = (\sigma_S, \sigma_R, CB)$

Predicting:

$$\begin{aligned} r_t &= \arg \min_{r \in R} E(s_t, r) \\ &= \arg \min_{r \in R} (\sigma_S, \sigma_R, CB [f(s_t, r)g) \end{aligned}$$

CoAT in terms of energy: Half Moon Running Example



Blue VS red class

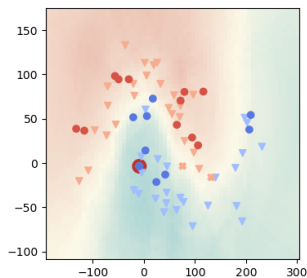
●HX: explained later

Energy map

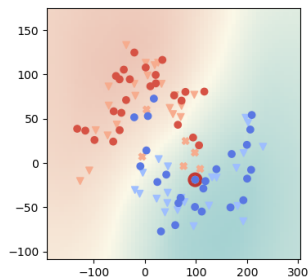
Case Base Competence

Competent CB: predicts the right outcome with high confidence

Less competent CB: wrong outcome or low confidence



Competent



Less competent

Case Base Competence

Energy-based pred. confidence¹:

$$\begin{array}{ll} \text{predicted outcome } r_t & \text{next best outcome } \hat{r} \\ E(s_t, r_t) & \min_{\hat{r} \notin r_t} E(s_t, \hat{r}) \end{array}$$

High confidence: $jE(s_t, r_t) - \min_{\hat{r} \notin r_t} E(s_t, \hat{r})j$ large

Wrong prediction: expected instead of predicted outcome

$$E(s_t, r_t) - \min_{\hat{r} \notin r_t} E(s_t, \hat{r}) > 0$$

$$(\) \quad \exists \hat{r} \notin r_t, E(s_t, \hat{r}) < E(s_t, r_t)$$

¹Yann LeCun et al. (2006). "A Tutorial on Energy-Based Learning". In: *Predicting Structured Data*. MIT Press

Case Base Competence

Training loss of energy-based models¹: prediction error of the CB
w.r.t. new case $c_t = (s_t, r_t)$

Minimum Classification Error (MCE) loss:

$$\ell_{MCE}(CB, c_t) = E(s_t, r_t) \quad \min_{\hat{r} \in r_t} E(s_t, \hat{r})$$

Hinge loss:

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

¹Yann LeCun et al. (2006). "A Tutorial on Energy-Based Learning". In: *Predicting Structured Data*. MIT Press

Case Base Competence

Energy-inspired CB competence, on reference set T :

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

Case Competence

Contribution of c to the CB competence:

$$C(c, CB, T) = C(CB, T) - C(CB \setminus fcg, T)$$

i.e., inversions involving case c

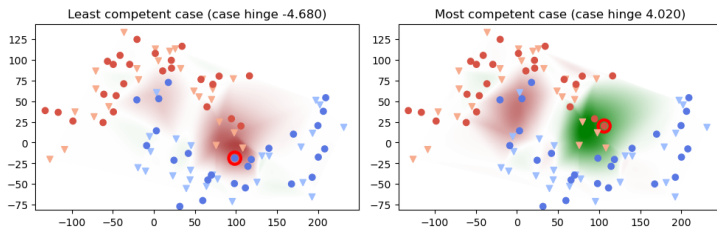
Case Influence & Expertise Area

Case competence *w.r.t.* a situation:

$$\text{influence}_{CB}(c, c_t) = \ell(CB, c_t) - \ell(CB \cap fcg, c_t)$$

$$C(c, CB, T) = \frac{1}{|T|} \sum_{c_t \in T} \text{influence}_{CB}(c, c_t)$$

Positive and negative influence of 2 cases (ℓ_{hinge}):



CB Maintenance: Case Deletion

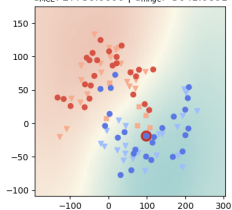
Algorithm 1 Case deletion procedure

Require: An initial case base CB and a reference set T
 while $|CB| > 0$ and stopping condition not fulfilled do
 $c_{worse} = \arg \min_{c \in CB} C(c, CB, T)$
 $CB = CB \setminus \{c_{worse}\}$
 end while

- $\sigma_S = \exp(-\frac{\|x - y\|^2}{2})$: Euclidean sim.
- $\sigma_R = 1$ if same class else 0: Class membership
- 10 CB 10 reference set (100 pairs), each 50 cases

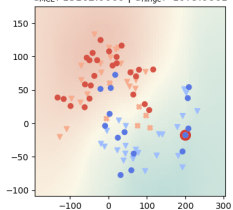
CB Maintenance: Case Deletion Example

$|CB| = 50$ | Macro F1 on ref.: 88.0%
 $CMCE: 27736.0000$ | $C_{Inge}: -3041.0601$

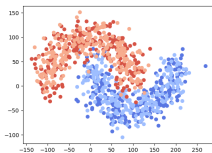
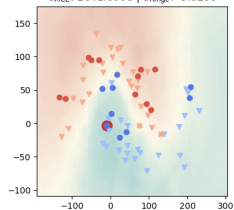


Half Moon

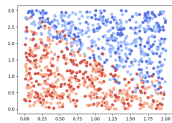
$|CB| = 40$ | Macro F1 on ref.: 88.0%
 $CMCE: 15102.0000$ | $C_{Inge}: -1079.0601$



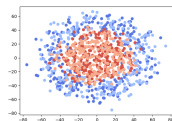
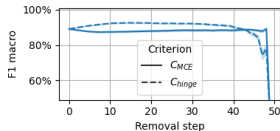
$|CB| = 20$ | Macro F1 on ref.: 96.0%
 $CMCE: 2041.0000$ | $C_{Inge}: -9.0200$



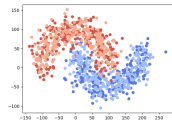
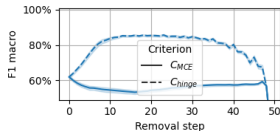
CB Maintenance: Performance and Competence



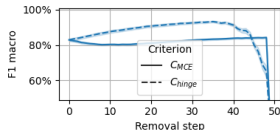
Line



Ring



Half Moon



Performance after removing (using C_{hinge}):

- 1 %: "noisy" case
- 2 !: "redundant" case
- 3 &: "useful" case

Main Findings

- Global approach VS usually local
- C_{hinge} more suitable than C_{MCE}
- C_{hinge} linked to CB performance, matches intuition of competence
- Can compress by 60% or 80% while increasing performance
- Robust *w.r.t.* initial CB & reference set
 - Properly distributed reference set δ bad bias
 - Enough initial cases in CB δ holes
- Appears to fit any distrib. (*c.f.* example Half Moon) even if σ_S not really suitable

(Some) Ongoing Work

- Competence linked with CB performance using other CBR methods than CoAT
- Real-world data
- More than 2 classes
- Dynamic aspect: impact of order of removal
- Robustness to errors, small differences, ... in similarities
- ...

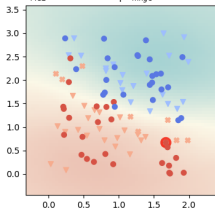
Questions

Thank you for your attention!



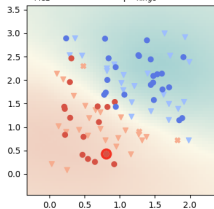
Case Deletion: Example

$|CB| = 50$ | Macro F1 on ref.: 84.0%
 $CMCE: 29430.0000$ | $Change: -2743.0801$

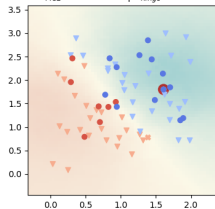


Line

$|CB| = 40$ | Macro F1 on ref.: 94.0%
 $CMCE: 21161.0000$ | $Change: -479.0300$



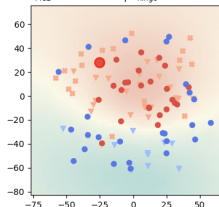
$|CB| = 20$ | Macro F1 on ref.: 98.0%
 $CMCE: 4425.0000$ | $Change: -1.0100$





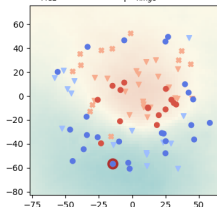
Case Deletion: Example

$|CB| = 50$ | Macro F1 on ref.: 65.3%
 $C_{MCE} = 11686.0000$ | $C_{Ninge} = -5263.1602$

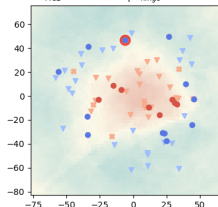


Ring

$|CB| = 40$ | Macro F1 on ref.: 75.8%
 $C_{MCE} = 5256.0000$ | $C_{Ninge} = -620.1200$



$|CB| = 20$ | Macro F1 on ref.: 89.9%
 $C_{MCE} = 1203.0000$ | $C_{Ninge} = -43.0500$





Case Deletion Run Time

Current implementation: quite optimized for CPU
Could go much faster with GPU

CB: 50 cases 1 case
50 reference cases

- Similarities computed once at the start $< 0.05s$
- Step 1: $2s$
- Step 25: $1s$
- Step 49: $< 0.1s$
- Total: $50s$

Bibliography

-  Badra, Fadi (2020). “A Dataset Complexity Measure for Analogical Transfer”. In: *Proc. of the 29th Int. Joint Conf. on Artificial Intelligence IJCAI*, pp. 1601–1607.
-  LeCun, Yann et al. (2006). “A Tutorial on Energy-Based Learning”. In: *Predicting Structured Data*. MIT Press.