Case-Based Prediction Using a Continuous Compatibility Measure

Chunyang Fan[1,2][a], Fadi Badra[2][b] and Marie-Jeanne Lesot[1][c]
[1]*LIP6, Sorbonne Université, CNRS, F-75005 Paris, France*
[2]*Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé - LIMICS, Université Sorbonne Paris Nord, INSERM, UMR 1142, F-93000, Bobigny, France*
firstname.lastname@lip6.fr, firstname.lastname@univ-paris13.fr

Keywords: Analogical Transfer, Similarity Measure Learning, Metric Learning, Case-Based Reasoning.

Abstract: Case-based prediction (CBP) addresses classification or regression tasks applying the analogical transfer principle (ATP) that infers information from similar known situations provided in the so-called case base: ATP leverages the assumption that similarity in some aspects, e.g. data features, implies similarity in others, e.g. data class. This paper considers the family of transfer by global optimization and complexity-based approaches. It proposes a continuous compatibility measure which quantifies the extent of ATP violation observed in the case base rather than merely counting the number of such violations. This new compatibility measure leads to a new classifier called CCoAT (Continuous Complexity-based Analogical Transfer). Furthermore, it allows for gradient-based optimization of the considered similarity measures: the paper proposes a metric learning method, called CCoSiL, to optimize the similarity measures for case-based prediction. The comparative experimental study conducted on several benchmark data sets shows the relevance of both the proposed classifier and metric learning method: CCoAT combined with CCoSiL achieves higher accuracy than when combined with the state-of-the-art metric learning method LMNN. In addition, CCoSiL can be applied to several CBP algorithms, it generalizes well to k-NN.

1 INTRODUCTION

Analogical transfer infers information from similar cases by assuming that their similarity in certain components implies similarity in others (Gust et al., 2008). Case-based prediction (CBP) methods (see e.g. (Gilboa and Schmeidler, 2010; Badra and Lesot, 2023) for surveys) apply this general principle to solve supervised machine learning tasks, classification or regression: they consider a data instance as a case with two components, respectively called situation and outcome, that correspond to the instance features and its label. An outcome similarity is then inferred from the situation similarity, to assess outcome candidates for a new situation through its comparison to situations with known outcomes, making the selection of the similarity measures a crucial choice. Due to the relations between similarity and metrics (Santini and Jain, 1999; Lesot et al., 2009), CBP thus raises issues related to the topic of metric learning (MeL) (Bellet et al., 2015; Suárez et al., 2021).

This paper considers the question of similarity learning in the case of CBP performed through transfer by global optimization, as proposed by the Complexity-based Analogical Transfer (CoAT) method (Badra, 2020). The latter relies on a global measure that captures the compatibility between situation similarities, outcomes similarities and the case base: for a new situation s, the predicted outcome is the candidate r that minimizes the value of the similarity incompatibility computed on the case base augmented with the candidate case (s,r). As detailed in Sec. 2, this global incompatibility measure has been shown to be related to loss functions used in MeL tasks (Badra et al., 2023), opening the way for a similarity learning approach optimized for CoAT. However, the incompatibility measure is piecewise-constant, preventing the application of gradient-based optimization methods.

In this paper, we first propose a new, continuous, compatibility measure to extend the CoAT method: while preserving its global semantics, it provides a continuous and more fine-grained view of incompat-

[a] https://orcid.org/0009-0003-9921-6353
[b] https://orcid.org/0000-0002-2437-8230
[c] https://orcid.org/0000-0002-3604-6647

ible configurations, quantifying their extent beyond the binary view taken by CoAT. We then propose to exploit the incompatibility continuity into a MeL optimization method to identify appropriate similarity measures.

After summarizing the context and related works in Sec. 2, we present the proposed continuous incompatibility measure and two induced classifiers in Sec. 3. Sec. 4 describes its optimization to learn the similarity measure, solving a MeL task. The experimental results of a study comparing the proposed CBP classifiers and MeL methods to benchmarks on reference datasets are discussed in Sec. 5. Finally, Sec. 6 concludes and discusses some directions for future works.

2 CONTEXT AND RELATED WORK

This section gives a brief overview of the two domains to which the proposed approach relates, before detailing the CBP algorithm on which the proposition relies, namely CoAT (Badra, 2020).

Case-Based Prediction. Case Based Prediction (CBP) applies the analogical transfer principle (ATP) to perform supervised learning tasks such as classification or regression, leveraging similarities between cases for outcome predictions. They can be categorized in 4 categories (Badra and Lesot, 2023), distinguishing between transfer by evidence support, continuity constraints, approximate reasoning or global optimization. All approaches crucially depend on the choice of the similarity measures used to quantify how alike cases are, and more precisely to compare situations on the one hand and outcomes on the other hand. Due to the relation between similarity measures and distances (Santini and Jain, 1999; Lesot et al., 2009), their selection relates to the topic of MeL.

Metric Learning. MeL aims at learning a distance function for machine learning tasks, including classification, clustering and information retrieval (Bellet et al., 2015; Suárez et al., 2021). A large variety of techniques have been proposed, among which one can mention linear, non-linear, local and histogram-based methods to name a few (see e.g. (Bellet et al., 2015) for a survey). A well-known example is the Large Margin Nearest Neighbor (LMNN) method (Weinberger et al., 2005), which learns the positive semi-definite matrix M that parametrizes a Mahalanobis distance d_M to optimize the performance of a k-NN classifier. Other approaches in case-based reasoning have been proposed e.g., learning weights for combining local measures in a data-driven approach (Jaiswal and Bach, 2019).

Deep metric learning employs neural networks to project data into embedding spaces, leading to pair-based methods using a contrastive loss (Hadsell et al., 2006), triplet-based methods (Hoffer and Ailon, 2015), which aim at bringing similar samples closer while pushing dissimilar samples apart, similarity cloud-based method (Gabel and Godehardt, 2015) and clustering-based methods capturing global data structures (Song et al., 2016). Additional approaches integrate attribute-specific measures, e.g., k-prototypes (Huang, 1998), or unsupervised hyperbolic methods for hierarchical data (Yan et al., 2021).

The CoAT Algorithm. The approaches in this paper are extensions of the Complexity-based Analogical Transfer algorithm (Badra, 2020), a CBP algorithm that performs transfer by global optimization and has been shown to bridge the gap to MeL (Badra et al., 2023).

CoAT considers a case base CB, i.e. a set of cases where each case c is a pair $(s,r) \in \mathcal{S} \times \mathcal{R}$, where $\mathcal{S}$ and $\mathcal{R}$ respectively denote the situation space and the outcome space. They are respectively equipped with the similarity measures $\sigma_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$ and $\sigma_{\mathcal{R}} : \mathcal{R} \times \mathcal{R} \to \mathbb{R}^+$.

Incompatibility Measure. CoAT relies on the analogical transfer principle (ATP): if two situations are more similar than two others, then their outcomes should also be more similar. It implements this principle by a global incompatibility indicator Γ: it counts the number of triplets (c_i, c_j, c_k) in CB that violate ATP, i.e. that are such that $\sigma_{\mathcal{S}}(s_i, s_j) \leq \sigma_{\mathcal{S}}(s_i, s_k)$ but $\sigma_{\mathcal{R}}(r_i, r_j) > \sigma_{\mathcal{R}}(r_i, r_k)$. Here, we denote $\theta := (\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, \text{CB})$, and let Ind_θ be the ATP violation indicator:

$$\text{Ind}_\theta(c_i, c_j, c_k) := \begin{cases} 1 & \text{if } \sigma_{\mathcal{S}}(s_i, s_j) \leq \sigma_{\mathcal{S}}(s_i, s_k), \\ & \sigma_{\mathcal{R}}(r_i, r_j) > \sigma_{\mathcal{R}}(r_i, r_k) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Γ can then be written as:

$$\Gamma\left(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, \text{CB}\right) = \sum_{(c_i,c_j,c_k) \in \text{CB}^3} \text{Ind}_\theta(c_i, c_j, c_k) \quad (2)$$

The key point of incompatibility is the inconsistency between the ordering of situation similarities and that of outcome similarities, but not the actual similarity values. It can thus be applied to extensions of similarity measures, e.g. a decreasing function of a distance, as $\sigma_{\mathcal{S}} = -d_M$.

CoAT Prediction. As a classifier or regressor, for a new situation s, CoAT predicts the outcome which minimizes Γ on the case base augmented with the new case (s,r):

$$\mathrm{CoAT}_\theta(s) := \arg\min_{r\in\mathcal{R}} \Gamma\left(\sigma_\mathcal{S}, \sigma_\mathcal{R}, \mathrm{CB}\cup\{(s,r)\}\right) \quad (3)$$

As shown by (Badra et al., 2023) that Γ can be interpreted within the framework of energy-based models (LeCun et al., 2006),

$$E_\theta^{\mathrm{CoAT}}(s,r) := \sum_{(c_i,c_j,c_k)\in(\mathrm{CB}\cup\{(s,r)\})^3} \mathrm{Ind}_\theta(c_i,c_j,c_k) \quad (4)$$

The algorithmic complexity of E_θ^{CoAT} is $O(|\,\mathrm{CB}\,|^3)$ evaluations of Ind_θ. In order to decrease it, (Badra et al., 2022) propose to predict the outcome that minimizes the following sum:

$$\sum_{(c_i,c_j,c_k)\in U(s,r)} \mathrm{Ind}_\theta(c_i,c_j,c_k) \quad (5)$$

where $U(s,r) := (\{(s,r)\}\times \mathrm{CB}^2)\cup(\mathrm{CB}\times\{(s,r)\}\times\mathrm{CB})\cup(\mathrm{CB}^2\times\{(s,r)\})$ is the set of triplets containing exactly one occurrence of the new case (s,r). The algorithmic complexity of prediction is then reduced to $3|\,\mathrm{CB}\,|^2$ computations of Ind_θ.

Beyond Prediction. (Badra et al., 2023) show that, beyond its use to infer the outcome for a new situation, the incompatibility measure Γ can be interpreted as an intrinsic indicator of the suitability of the similarity measures to the case base.

For prediction tasks, especially classification, $\sigma_\mathcal{R}$ is often fixed. E_θ^{CoAT} can thus represent the suitability of the similarity measure $\sigma_\mathcal{S}$ w.r.t. CB. This perspective motivates us to study a MeL approach to optimize $\sigma_\mathcal{S}$ with E_θ^{CoAT}.

3 PROPOSED CONTINUOUS INCOMPATIBILITY MEASURE

This section presents the continuous incompatibility measure we propose, motivated by the discrete nature of the Γ indicator used in CoAT that implies its discontinuity and the impossibility to use traditional gradient-based optimization methods. As a consequence, although CoAT constitutes an energy-based model, its energy function E_θ^{CoAT} cannot be used to define a loss function to optimize the similarity measures. We then propose the continuous incompatibility measure, focusing on the classification case.

3.1 Motivation

We first analyze the building block of CoAT, the ATP violation measure Ind_θ. By definition (Eq. 1), Ind_θ can only take two values, 0 or 1. Thus, as a sum of Ind_θ (Eq. 4), E_θ^{CoAT} is a step function w.r.t. (c_i,c_j,c_k) and θ.

This property is illustrated in Fig.1 where $\mathcal{S}=\mathbb{R}$, $\mathcal{R}=\{\bullet,\blacktriangle\}$ and CB only contains one case of each class: $\mathrm{CB}=\{\alpha,\beta\}$ with $\alpha:=(-1,\bullet)$ and $\beta:=(1,\blacktriangle)$. We consider that $\sigma_\mathcal{S}$ is the opposite of the Euclidean distance, denoted $\sigma^*_\mathcal{S}(s_i,s_j)=-d_E(s_i,s_j)$, and the classical classification choice $\sigma_\mathcal{R}(r,r')=\mathbb{1}\{r=r'\}$.

For case $c=(s,\bullet)$, let us denote $f(c):=\mathrm{Ind}_\theta(c,\alpha,\beta)$. It holds that, if $s\le 0$ (e.g. $s_1=-0.25$ in Fig.1), $f(c)=0$ and if $s>0$ (e.g. $s_2=0.25$ and $s_3=0.75$ in Fig.1), $f(c)=1$. Indeed, when $s>0$, $\sigma^*_\mathcal{S}(s,\alpha)<\sigma^*_\mathcal{S}(s,\beta)$ while $\sigma_\mathcal{R}(\bullet,\bullet)>\sigma_\mathcal{R}(\bullet,\blacktriangle)$. Hence, the similarity order between the situation space and the outcome space is reversed, whereas is it identical for $s\le 0$. This makes Ind_θ jump from 0 to 1 when s passes 0. Besides, it is constant for $s\le 0$ and $s>0$, making its derivative zero almost everywhere (a.e.).

This variation of the relative positions of the 3 situations can be transferred to a variation of $\sigma_\mathcal{S}$. For example, $\sigma_\mathcal{S}=-d_M$ is equivalent to $-d_E$ after a linear transformation of the original space. Thus, a variation of $\sigma_\mathcal{S}$ is equivalent to that of relative positions of (s_i,s_j,s_k) after a linear transformation : denoting $M=LL^\top$ by Cholesky decomposition, it holds that $\sigma_\mathcal{S}(s_i,s_j)-\sigma_\mathcal{S}(s_i,s_k)=\sigma^*_\mathcal{S}(L^\top s_i,L^\top s_j)-\sigma^*_\mathcal{S}(L^\top s_i,L^\top s_k)$. The same type of interpretation can be applied for any $\sigma_\mathcal{S}$ continuous w.r.t. s_i and s_j.

Consequently, E_θ^{CoAT} is discrete and its gradient is 0 a.e. w.r.t. θ. Thus, traditional gradient-based optimization methods cannot be applied to learn similarity measures by optimizing E_θ^{CoAT}.

In addition, Ind_θ only indicates whether a triplet violates ATP, regardless of its level. As showed in Fig.1, for $c=(s,\bullet)$ on the positive side, even if s is very close to 0, $\mathrm{Ind}_\theta(c,\alpha,\beta)=1$ just as when s is far away from 0. As a result, E_θ^{CoAT} is not sensitive to the level of violation.

3.2 New ATP Violation Measure

To address these issues, we propose a new ATP violation measure Ext_θ: it does not qualitatively determine whether ATP is violated, but quantitatively assesses

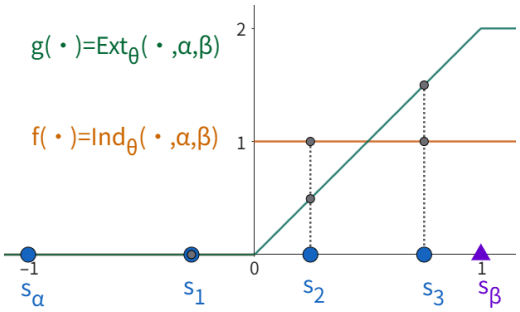

Figure 1: An example of Ind_θ and Ext_θ where CB $:= \{\alpha := (-1, \bullet), \beta := (1, \blacktriangle)\}$, $c_1 := (-0.25, \bullet)$, $c_2 := (0.25, \bullet)$, $c_3 := (0.75, \bullet)$

its extent:

$$\text{Ext}_\theta(c_i, c_j, c_k) \tag{6}$$
$$= \begin{cases} \sigma_S(s_i, s_k) - \sigma_S(s_i, s_j) & \text{if } \sigma_S(s_i, s_j) \leq \sigma_S(s_i, s_k), \\ & \sigma_\mathcal{R}(r_i, r_j) > \sigma_\mathcal{R}(r_i, r_k) \\ 0 & \text{otherwise} \end{cases}$$

As Ind_θ, Ext_θ returns 0 when the triplet does not violate ATP; however, when it does, Ext_θ does not jump to 1 but quantifies the extent of violation by the difference $\sigma_S(s_i, s_k) - \sigma_S(s_i, s_j)$. This property is illustrated in Fig.1 by the variation of $g(c) := \text{Ext}_\theta(c, \alpha, \beta)$. When s passes 0 from left to right, $g(c)$ starts to increase from 0, and $g(c) = 2s$ when $0 \leq s \leq 1$: the farther s moves away from 0, the higher Ext_θ becomes.

Another important advantage of Ext_θ is its continuity w.r.t. θ when σ_S is continuous. Indeed, we can rewrite Ext_θ as a product of a continuous function and a constant function w.r.t. σ_S:

$$\text{Ext}_\theta(c_i, c_j, c_k) = \max(0, \sigma_S(s_i, s_k) - \sigma_S(s_i, s_j)) \cdot \mathbb{1}\{\sigma_\mathcal{R}(r_i, r_j) > \sigma_\mathcal{R}(r_i, r_k)\} \tag{7}$$

where $\mathbb{1}\{\sigma_\mathcal{R}(r_i, r_j) > \sigma_\mathcal{R}(r_i, r_k)\}$ is independent of σ_S. For computational optimization, it generally suffices that Ext_θ is continuous with non-zero gradients in most parts of the parameter space. Moreover, Ext_θ is differentiable a.e. w.r.t. θ when σ_S is differentiable (e.g., d_M or Gaussian kernel).

3.3 Proposed CCoAT Classifiers

Based on the proposed ATP violation measure Ext_θ, we propose two variants of CoAT's energy function and define the corresponding classifiers.

Energy Function with Violation Extent. One straightforward way to define a new variant of E_θ^{CoAT} is to replace Ind_θ by Ext_θ in Eq. 4 and sum over all triplets in the case base augmented with the new case (s, r):

$$E_\theta^{\text{A}}(s, r) := \sum_{(c_i, c_j, c_k) \in U(s,r)} \text{Ext}_\theta(c_i, c_j, c_k) \tag{8}$$

where A stands for "All Triplets". As the sum of Ext_θ over $U(s, r)$, E_θ^{A} can be interpreted as the total extent of ATP violation. E_θ^{A} inherits Ext_θ sensitivity of violation and mathematical properties such as continuity and (a.e.) differentiability. Thus, when σ_S is differentiable, E_θ^{A} can be optimized by using gradient-based optimization methods, which is exploited in Sec. 4.

Energy Function with Violation Extent and Reduced Triplet Set. To reduce the algorithmic complexity and to adapt to imbalanced classes, we propose E_θ^{R} with three characteristics, commented in turn below: not all triplets are considered, not all permutations of each triplet are considered, and a normalization factor is introduced. Formally,

$$E_\theta^{\text{R}}(s, r) := \frac{1}{|T(r)|} \sum_{(c_j, c_k) \in T(r)} \text{Ext}_\theta((s, r), c_j, c_k)$$

where $T(r) := \{(c_j, c_k) \in \text{CB}^2 \mid r_j = r, r_k \neq r\}$ is the Reduced triplet set, for which R stands.

First, not all triplets are taken into account: some of them can be considered as trivial, insofar as they are not useful to compute the compatibility indicator, because they do not violate ATP, whatever σ_S is. Such trivial triplets for instance include the ones containing 3 cases from a same class (i.e. $r_i = r_j = r_k$): they do not satisfy the condition $\sigma_\mathcal{R}(r_i, r_j) > \sigma_\mathcal{R}(r_i, r_k)$ and lead to a 0 contribution in the energy computation. Likewise, for multi-class tasks, triplets containing cases from 3 different classes are trivial. As a consequence, only triplets satisfying $\sigma_\mathcal{R}(r_i, r_j) > \sigma_\mathcal{R}(r_i, r_k)$ are needed, which is equivalent to $r_j = r_i$ and $r_k \neq r_i$. This justifies replacing the set $U(s, r)$ with the reduced set $T(r)$. With this choice, the condition $\sigma_\mathcal{R}(r_i, r_j) > \sigma_\mathcal{R}(r_i, r_k)$ in Ext_θ (Eq. 7) is always satisfied, i.e. $T(r)$ contains no trivial triplet.

Second, not all permutations are taken into account: as CoAT, E_θ^{A} uses each triplet 6 times, leading to some unnecessary computations. Indeed, without loss of generality, suppose that c_i, c_j, c_k come from 2 classes, such that $r_i = r_j$ and $\text{Ext}_\theta(c_i, c_j, c_k) \neq 0$. Then $\text{Ext}_\theta(c_i, c_k, c_j) = 0$ regardless of σ_S: since $\sigma_\mathcal{R}(r_i, r_k) > \sigma_\mathcal{R}(r_i, r_j)$ cannot be satisfied simultaneously with $\sigma_\mathcal{R}(r_i, r_j) > \sigma_\mathcal{R}(r_i, r_k)$. Consequently, one of these permutations is trivial. Moreover, we consider the following approximation: $\text{Ext}_\theta(c_j, c_i, c_k)$ is unlikely to be largely different from $\text{Ext}_\theta(c_i, c_j, c_k)$ if σ_S is a "suitable" similarity measure. It is thus not necessary to consider both permutations. We thus

propose to take into account only two orderings for each triplet, imposing that (s,r) appears in the first position and considering $((s,r),c_j,c_k)$.

Reducing the considered triplets and permutations leads to the decrease of computational complexity in E^{R}_θ as compared to E^{A}_θ (which already relies on the reduced set $U(s,r)$): let us denote $N := |\mathcal{R}|$ the number of classes and $\overline{n} := |\,\mathrm{CB}\,|/N$ the average number of instances per class. It holds that $|T(r)| = \overline{n} \times (N-1)\overline{n}$ on average, thus,

$$\frac{|T(r)|}{|U(s,r)|} = \frac{(N-1)\overline{n}^2}{3(N\overline{n})^2} = \frac{N-1}{3N^2},$$

which means that E^{R}_θ considers only $\frac{N-1}{3N^2}$ of the triplets in E^{A}_θ on average. For $N=2$, it reduces 91.67% of triplets to evaluate on average.

The third characteristic of the proposed E^{R}_θ energy function is the normalizing factor $1/|T(r)|$, which allows to avoid undesired behaviors in the case of imbalanced classes. Indeed, otherwise, minority classes may be overwhelmed by majority classes, which may lead to bias in optimizing of the similarity measure. The explicit control of class brought by $T(r)$ endows Ext_θ with a notion of average extent of violation for class r, which helps avoid bias from class imbalance.

Classifiers. Both E^{A}_θ and E^{R}_θ constitute continuous energy functions that can be used to define new classifiers:

$$\mathrm{CCoAT}^{\mathrm{A}}_\theta(s) := \arg\min_{r\in\mathcal{R}} E^{\mathrm{A}}_\theta(s,r) \tag{9}$$

$$\mathrm{CCoAT}^{\mathrm{R}}_\theta(s) := \arg\min_{r\in\mathcal{R}} E^{\mathrm{R}}_\theta(s,r) \tag{10}$$

Both classifiers inherit the idea of analogical transfer from CoAT, but they use a continuous and more sensitive energy function to measure the compatibility between the similarity measures and the case base.

4 PROPOSED MeL METHOD

This section exploits the proposed energy function to design MeL algorithms for CBP, called CCoSiL that stands for Continuous Complexity-based Similarity Learning.

4.1 Proposed Loss Functions

The first loss function combines the proposed energy function with MCE (Minimum Classification Error) (Juang et al., 1997) defined as

$$\ell_{\mathrm{MCE}}(s,r;E_\theta) := \mathrm{Sig}\left(E_\theta(s,r) - \min_{r'\in\mathcal{R}\setminus\{r\}} E_\theta(s,r')\right)$$

In this paper, we set $\mathrm{Sig}(x) := \frac{1}{1+e^{-x}}$. The main motivation for using the sigmoid function is that minimizing the induced loss function is related to minimizing the probability that the classifier predicts the wrong class (Juang and Katagiri, 1992).

However, the MCE loss is sensitive to outliers. In particular, when the gap between the energy of the correct class and that of others is large, due to the exponential term, ℓ_{MCE} can stay at extremes 0 or 1 during numerical optimization, leading to a very small gradient and thus a slow convergence.

The second loss function is the hinge loss (LeCun and Huang, 2005), defined as

$$\ell_{\mathrm{H}}(s,r;E_\theta) := \max\left(0, \lambda + E_\theta(s,r) - \min_{r'\in\mathcal{R}\setminus\{r\}} E_\theta(s,r')\right)$$

The hinge loss penalizes cases where the energy of the correct class is not lower than that of other classes by at least a margin λ. Even if it is not differentiable at some points, it is sub-differentiable, which makes it amenable to optimization with sub-gradient methods.

4.2 Optimization of Similarity Measures

To optimize the similarity measures, we minimize the total loss on the training set $\mathcal{D} = \{(s_i,r_i)\}_{i=1}^n$:

$$\widehat{\theta} := \mathrm{CCoSiL}(E_\theta,\ell) := \arg\min_\theta \frac{1}{n}\sum_{i=1}^n \ell(s_i,r_i;E_\theta) \tag{11}$$

where E_θ is either E^{A}_θ or E^{R}_θ, and ℓ is either ℓ_{MCE} or ℓ_{H}. This leads to 4 possible configurations for CCoSiL, respectively denoted $\mathrm{CCoSiL}^{\mathrm{A}}_{\mathrm{MCE}}$, $\mathrm{CCoSiL}^{\mathrm{A}}_{\mathrm{H}}$, $\mathrm{CCoSiL}^{\mathrm{R}}_{\mathrm{MCE}}$ and $\mathrm{CCoSiL}^{\mathrm{R}}_{\mathrm{H}}$.

As earlier mentioned, in practice, $\sigma_{\mathcal{R}}$ and CB are often fixed, only $\sigma_{\mathcal{S}}$ needs to be optimized, which most often translates to optimizing its parameters, e.g. the matrix M for the Mahalanobis distance d_M. Therefore, the optimization problem in Eq. 11 can be considered as a parametrized estimation problem, and solved by gradient-based optimization methods, such as Adam.

5 EXPERIMENTAL STUDY

This section presents the experimental study conducted to assess the relevance of the 2 proposed classifiers $\mathrm{CCoAT}^{\mathrm{A}}_\theta$ and $\mathrm{CCoAT}^{\mathrm{R}}_\theta$ and the 4 proposed MeL methods.

5.1 Protocol

We consider six classical UCI benchmarks with continuous features: Iris, User, Breast Cancer, PIMA, Wine, and Seeds. We divide each dataset into a training set $\mathcal{D}$ and a test set $\mathcal{T}$. $\mathcal{D}$ is used to train the similarity measures and as CB for the classifiers (C)CoAT (i.e., CB $= \mathcal{D}$), while $\mathcal{T}$ is used to evaluate their accuracy. After splitting each dataset, we normalize the features by average and standard deviation of $\mathcal{D}$. To mitigate selection bias, we employ 10-fold cross-validation.

As for the CBP classifiers, we compare prediction accuracies of CoAT_θ, $\text{CCoAT}_\theta^\text{A}$, $\text{CCoAT}_\theta^\text{R}$ and the classical k-NN classifier. All classifiers are equipped with the same similarity measure $\sigma_\mathcal{S} = -d_M^2$, whose parameter M is learned.

Regarding the MeL methods, we compare 3 configurations: (i) no training, i.e. setting M to the identity matrix, in which case the Mahalanobis distance corresponds to the Euclidean distance; (ii) the LMNN method (Weinberger et al., 2005), dedicated to optimizing the Mahalanobis distance for k-NN; (iii) the proposed method CCoSiL with all 4 configurations. For LMNN, the learning rate is set to 10^{-6} with automatic epoch stopping (50 - 1000 epochs). The k value of k-NN is set to 3, the default value of LMNN. For CCoSiL, we use Adam optimizer with learning rate 10^{-2} and 50 epochs. For ℓ_H, the margin λ is set to 1.

5.2 Results and Analysis

The accuracy results[1] are shown in Table 1. The detailed analysis first studies, for the complexity based CBP classifiers (C)CoAT, the effectiveness of the proposed MeL method CCoSiL: it compares (C)CoAT + CCoSiL to (C)CoAT without training (i.e. the identity matrix) and (C)CoAT + LMNN. Second, it compares the proposed combination of (C)CoAT + CCoSiL to the reference CBP combination k-NN + LMNN. The third step compares the generalizability of the MeL methods CCoSiL and LMNN to classifiers they were not specifically designed for, examining the combinations (C)CoAT + LMNN and k-NN + CCoSiL.

Effectiveness of CCoSiL **for (C)CoAT Classifiers.** Table 1 shows that, except for the configuration $\text{CCoSiL}_\text{MCE}^\text{A}$, all configurations of CCoSiL significantly improve the accuracy of both CoAT and CCoAT classifiers on almost all datasets, with gains from 8.81 to 10.40 percentage points (ppt) on average. These results confirm the effectiveness of the proposed MeL method.

Specifically, among all configurations of CCoSiL, we observe the following *best configurations* for (C)CoAT, with average accuracies gains:

- $\text{CCoAT}_\theta^\text{R}$: $\text{CCoSiL}_\text{H}^\text{R}$ improves 10.11 ppt; $\text{CCoSiL}_\text{MCE}^\text{R}$ improves 9.89 ppt.
- $\text{CCoAT}_\theta^\text{A}$: $\text{CCoSiL}_\text{MCE}^\text{R}$ improves 10.40 ppt; $\text{CCoSiL}_\text{H}^\text{A}$ improves 10.22 ppt.
- CoAT_θ: $\text{CCoSiL}_\text{MCE}^\text{R}$ improves 10.37 ppt; $\text{CCoSiL}_\text{H}^\text{A}$ improves 10.13 ppt.

Our investigation into the lack of improvement for $\text{CCoSiL}_\text{MCE}^\text{A}$ reveals that, as discussed in Sec. 4.1, ℓ_MCE stays at the extremes during training, which leads to no decrease direction for optimization.

Furthermore, we observe that, across all MeL methods, $\text{CCoAT}_\theta^\text{R}$ can achieve the same accuracy as $\text{CCoAT}_\theta^\text{A}$, and higher than CoAT_θ. CCoSiL using E_θ^R can also achieve the same or better accuracy than that using E_θ^A. This confirms the effectiveness of the proposed energy function E_θ^R, which allows both classifier and MeL method to achieve better performance with lower algorithmic complexity.

Comparison of CCoSiL **and LMNN for (C)CoAT Classifiers.** With the best configurations of CCoSiL, (C)CoAT classifiers achieve better average accuracies than when using LMNN, with average gains from 2.44 to 2.76 ppt. Besides, their best accuracies are achieved by the best CCoSiL configurations on 5/6 datasets, except for Wine. Therefore, the proposed MeL method CCoSiL appears more effective than LMNN for (C)CoAT classifiers.

Comparison of (C)CoAT with CCoSiL **and k-NN with LMNN.** We then compare the performance of (C)CoAT classifiers with CCoSiL to the performance of the reference CBP classifier, namely k-NN with LMNN. From Table 1 we observe that, with the best configurations of CCoSiL, (C)CoAT classifiers achieve slightly better performance on 4 or 5 out of 6 datasets, as compared to k-NN with LMNN: the average accuracies show gains from 0.39 to 0.77 ppt. Therefore, the proposed (C)CoAT classifiers with the proposed MeL method CCoSiL are comparable to or slightly better than k-NN with LMNN.

Note that, however, whereas LMNN is specifically designed for learning the parameter of the Mahalanobis distance, CCoSiL is general and can be applied to train any differentiable similarity measures. In that sense, CCoSiL is more flexible than LMNN w.r.t. the metric family.

[1] The source code is available at https://gitlab.lip6.fr/chunyang/ccosil-icaart2026

Table 1: Prediction accuracies of the combinations of the 4 considered classifiers and the 6 considered MeL methods on 6 datasets (in %). The best accuracy for each dataset is underlined, and the best accuracy for each classifier and each dataset among all MeL methods is in bold.

Classifier	Metric	Iris	User	Breast	PIMA	Wine	Seeds	average
k-NN	Identity	94.00%	83.12%	**97.01%**	73.83%	95.56%	91.90%	89.24%
	LMNN	95.33%	92.55%	96.48%	**74.10%**	**99.44%**	92.86%	**91.79%**
	CCoSiL$^{\mathrm{A}}_{\mathrm{MCE}}$	95.33%	83.12%	**97.01%**	73.83%	95.56%	91.90%	89.46%
	CCoSiL$^{\mathrm{A}}_{\mathrm{H}}$	**96.67%**	94.52%	95.26%	72.67%	97.78%	**93.33%**	91.71%
	CCoSiL$^{\mathrm{R}}_{\mathrm{MCE}}$	93.33%	**95.29%**	96.48%	71.10%	97.22%	92.38%	90.97%
	CCoSiL$^{\mathrm{R}}_{\mathrm{H}}$	94.67%	93.79%	95.96%	73.83%	97.22%	**93.33%**	91.47%
CoAT$_\theta$	Identity	87.33%	55.59%	92.97%	72.66%	92.78%	90.95%	82.05%
	LMNN	95.33%	87.32%	93.85%	71.49%	**99.44%**	92.38%	89.97%
	CCoSiL$^{\mathrm{A}}_{\mathrm{MCE}}$	89.33%	55.59%	92.97%	72.66%	92.78%	90.95%	82.38%
	CCoSiL$^{\mathrm{A}}_{\mathrm{H}}$	**96.67%**	94.03%	95.25%	**76.69%**	96.63%	**93.81%**	92.18%
	CCoSiL$^{\mathrm{R}}_{\mathrm{MCE}}$	95.33%	**95.76%**	**96.31%**	76.56%	97.19%	93.33%	**92.41%**
	CCoSiL$^{\mathrm{R}}_{\mathrm{H}}$	94.67%	95.04%	93.85%	71.36%	97.19%	93.33%	90.91%
CCoAT$^{\mathrm{A}}_\theta$	Identity	85.33%	59.31%	91.74%	72.79%	92.78%	90.95%	82.15%
	LMNN	94.67%	89.55%	93.32%	71.87%	**99.44%**	91.43%	90.05%
	CCoSiL$^{\mathrm{A}}_{\mathrm{MCE}}$	86.67%	59.31%	91.74%	72.79%	92.78%	90.95%	82.37%
	CCoSiL$^{\mathrm{A}}_{\mathrm{H}}$	95.33%	95.02%	95.78%	**77.08%**	97.19%	**93.81%**	92.37%
	CCoSiL$^{\mathrm{R}}_{\mathrm{MCE}}$	**96.00%**	**96.01%**	**96.31%**	76.43%	97.19%	93.33%	**92.55%**
	CCoSiL$^{\mathrm{R}}_{\mathrm{H}}$	**96.00%**	**96.01%**	92.62%	71.62%	96.63%	92.86%	90.96%
CCoAT$^{\mathrm{R}}_\theta$	Identity	84.67%	62.24%	89.80%	75.13%	90.52%	92.38%	82.46%
	LMNN	94.00%	87.06%	92.44%	74.48%	**99.44%**	91.43%	89.81%
	CCoSiL$^{\mathrm{A}}_{\mathrm{MCE}}$	86.00%	62.24%	89.80%	75.13%	90.52%	92.38%	82.68%
	CCoSiL$^{\mathrm{A}}_{\mathrm{H}}$	94.67%	94.52%	95.43%	75.78%	97.19%	**92.86%**	91.74%
	CCoSiL$^{\mathrm{R}}_{\mathrm{MCE}}$	95.33%	**95.03%**	96.13%	76.95%	97.75%	**92.86%**	92.34%
	CCoSiL$^{\mathrm{R}}_{\mathrm{H}}$	**96.00%**	**95.03%**	**96.66%**	**77.09%**	97.75%	**92.86%**	**92.57%**

Generalizability of CCoSiL**.** We finally study the flexibility of the MeL methods w.r.t. the CBP classifier they are applied to: we investigate their applicability to non-dedicated classifiers, applying CCoSiL for k-NN and LMNN for (C)CoAT.

For k-NN, LMNN improves its accuracy on 5/6 datasets, with an average gain of 2.56 ppt. CCoSiL$^{\mathrm{A}}_{\mathrm{H}}$ brings higher improvement on 3/6 datasets, but only 0.09 ppt lower improvement on average. Therefore, we can conclude that CCoSiL is also effective for the k-NN classifier, and is comparable to LMNN.

For (C)CoAT, LMNN improves their accuracies on 4 or 5 out of 6 datasets, with average gains from 7.35 to 7.92 ppt. Compared to LMNN, the best configurations of CCoSiL bring higher improvement on 5 datasets, and 2.21 to 2.76 ppt higher improvement on average. Therefore, we can conclude that while LMNN is also effective for (C)CoAT classifiers, CCoSiL is more effective.

6 CONCLUSION

In this paper, we introduced a refined, quantified, measure of violation of the Analogical Transfer Principle and two induced continuous energy functions, E^{A}_θ and E^{R}_θ. They lead to the proposition of continuous variants for complexity based analogical transfer: CCoAT$^{\mathrm{A}}_\theta$ and CCoAT$^{\mathrm{R}}_\theta$. In addition, exploiting the fact that this continuity enables gradient-based optimization, we propose a general metric learning framework CCoSiL.

The obtained experimental results show the relevance of CCoSiL that consistently allows to boost (C)CoAT accuracy and efficiency. E^{R}_θ delivers the best overall performance with significantly reduced algorithmic complexity. CCoSiL is more effective than LMNN for (C)CoAT and achieves performance comparable to or slightly better than k-NN with LMNN, while being more flexible as it applies to di-

verse similarity measures. This work illustrates how tailored metric learning can strengthen CBP classifiers and lessen reliance on expert-chosen similarities, while remaining transferable to other CBP methods.

Ongoing works include further experimental studies, in particular investigating other families of parametrized similarity measures to study their impact on (C)CoAT classifiers and the effectiveness of CCoSiL in optimizing them. Directions for future works include the exploration of flexible variants of Ext_θ, in particular through a hyperparameter to adjust the tolerance to ATP violations: one may consider that below a given parametrized threshold, they may be negligible and should not impact the optimization process, where this parameter may for instance be adjusted to the characteristic of the processed data set.

ACKNOWLEDGEMENTS

This work is funded by the SMeLT project, ANR-22-CE23-0032-03.

REFERENCES

Badra, F. (2020). A Dataset Complexity Measure for Analogical Transfer. In *IJCAI'20*.

Badra, F. and Lesot, M.-J. (2023). Case-based prediction – A survey. *IJAR*.

Badra, F., Lesot, M.-J., Barakat, A., and Marsala, C. (2022). Theoretical and Experimental Study of a Complexity Measure for Analogical Transfer. In *ICCBR'22*.

Badra, F., Lesot, M.-J., Marquer, E., and Couceiro, M. (2023). Some Perspectives on Similarity Learning for Case-Based Reasoning and Analogical Transfer. In *IARML@IJCAI'23*.

Bellet, A., Habrard, A., and Sebban, M. (2015). *Metric Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer.

Gabel, T. and Godehardt, E. (2015). Top-Down Induction of Similarity Measures Using Similarity Clouds. In *ICCBR'15*.

Gilboa, I. and Schmeidler, D. (2010). Case based Predictions: Introduction. *Introduction to Case-Based Prediction. World Scientific Publishers*.

Gust, H., Krumnack, U., Kühnberger, K.-U., and Schwering, A. (2008). Analogical Reasoning: A Core of Cognition. *KI*.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality Reduction by Learning an Invariant Mapping. In *CVPR'06*.

Hoffer, E. and Ailon, N. (2015). Deep Metric Learning Using Triplet Network. In *Similarity-Based Pattern Recognition*.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *DMKD'98*.

Jaiswal, A. and Bach, K. (2019). A Data-Driven Approach for Determining Weights in Global Similarity Functions. In *ICCBR'19*.

Juang, B.-H., Hou, W., and Lee, C.-H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech Audio Process.*

Juang, B.-H. and Katagiri, S. (1992). Discriminative learning for minimum error classification (pattern recognition). *IEEE Trans. Signal Process.*

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. J. (2006). *A Tutorial on Energy-Based Learning*. MIT press.

LeCun, Y. and Huang, F. J. (2005). Loss functions for discriminative training of energy-based models. In *AISTATS'05*.

Lesot, M.-J., Rifqi, M., and Benhadda, H. (2009). Similarity measures for binary and numerical data: a survey. *IJKESDP*.

Santini, S. and Jain, R. (1999). Similarity measures. *IEEE Trans. Pattern Anal. Mach. Intell.*

Song, H. O., Xiang, Y., Jegelka, S., and Savarese, S. (2016). Deep Metric Learning via Lifted Structured Feature Embedding. *CVPR'16*.

Suárez, J. L., García, S., and Herrera, F. (2021). A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*.

Weinberger, K. Q., Blitzer, J., and Saul, L. (2005). Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *NIPS'05*.

Yan, J., Luo, L., Deng, C., and Huang, H. (2021). Unsupervised Hyperbolic Metric Learning. In *CVPR'21*.