Hierarchical Learning of Dependent Concepts for Human Activity Recognition

Anonymous Author
Anonymous Institution

Abstract

In multi-class classification tasks, like human activity recognition, it is often assumed that classes are separable. In real applications, this assumption becomes strong and generates inconsistencies. Besides, the most commonly used approach is to learn classes one-by-one against the others. This computational simplification principle introduces strong inductive biases on the learned theories. In fact, the natural connections among some classes, and not others, deserve to be taken into account. In this paper, we show that the organization of overlapping classes (multiple inheritances) into hierarchies considerably improves classification performances. This is particularly true in the case of activity recognition tasks featured in the Sussex-Huawei Locomotion (SHL) dataset. After theoretically showing the exponential complexity of possible class hierarchies, we propose an approach based on transfer affinity among the classes to determine an optimal hierarchy for the learning process. Extensive experiments on the SHL dataset show that our approach improves classification performances while reducing the number of examples needed to learn.

1 Introduction

Many real-world applications considered in machine learning exhibit dependencies among the various concepts (or classes) which need to be learned [7, 24, 20]. This is particularly the case in human activity recognition from wearable sensor deployments which constitutes the main focus of our paper. Indeed, are the criteria used to distinguish between the activities (concepts) *running* and *walking* the same as those used to distinguish between *driving a car* and *being in a bus*? what about distinguishing each individual activity against the remaining ones taken as a whole? Similarly, during the annotation process, when should someone consider that *walking* at a higher pace corresponds actually to *running*? These questions naturally arise in the case of the Sussex-Huawei Locomotion (SHL) dataset [8] which exhibits such dependencies. The considered activities in this dataset are difficult to separate due to the existence of many overlaps among certain activities. Some of the important causes for these overlaps are: (1) the on-body sensor deployment featured by this dataset, due to sensors coverage overlaps, tend to capture movements that are not necessarily related to a unique activity. Authors in [10], for example, have exhibited such overlaps. (2) The difficulty of data annotation during data collection conducted in real-world conditions. For instance, the annotation issues can include the time-shift of a label with respect to the activity [22], as well as wrong or missing labels [15]. Similarly, long lines of research in computer vision [17, 23] and time-series analysis [22, 9, 15] raised these issues which hinder the development and large-scale adoption of these applications.

To solve these problems, we propose an original approach for structuring the considered concepts into hierarchies in a way that very similar concepts are grouped together and tackled by specialized classifiers. The idea is that classifications at different levels of the hierarchy may rely on different features, or different combinations of the same features [34]. Indeed, many real-world classification problems are naturally cast as hierarchical classification problems [2, 34, 28, 29]. A work on the semantic relationships among the categories in a hierarchical structure shows that they are usually of the type *generalization-specialization* [34]. In other words, the lower-level categories are supposed to have the same general properties as the higher-level categories plus additional more specific proper-

Preliminary work. Under review by AISTATS 2021. Do not distribute.

ties. The problem at hand is twice difficult as we have to, first, find the most appropriate hierarchical structure and, second, find optimal learners assigned to the nodes of the hierarchical structure.

We propose a data-driven approach to structure the considered concepts in a bottom-up approach. We start by computing the affinities and dependencies that exist among the concepts and fuse hierarchically the closest concepts with each other. We leverage for this a powerful technique based on transfer which showed interesting empirical properties in various domains [31, 16]. Taking a bottom-up approach allows us to leverage learning the complete hierarchy (including the classifiers assigned to each non-leaf node) incrementally by reusing what was learned on the way. Our contributions are as follows: (1) we propose a theoretical calculation for computing the total number of tree hierarchical combinations (the search space for the optimal solution) based on the given number of concepts; (2) we propose an approach based on transfer affinity to determine an optimal organization of the concepts that improves both learning performances and accelerates the learning process; (3) Extensive experiments show the effectiveness of organizing the learning process. We noticeably get a substantial improvement of recognition performances over a baseline which uses a flat classification setting; (4) we perform a comprehensive comparative analysis of the various stages of our proposed approach which raises interesting questions about concept dependencies, the generalization of the final classifiers, and the amount of supervision.

2 Problem Statement

In this section, we briefly review the problem of hierarchical structuring of concepts in terms of formulation and background. We then provide a complexity analysis of the problem size and its search space.

2.1 Problem Formulation and Background

Let $\mathcal{X} \subset \mathbb{R}^n$ be the inputs vector [1] and let $\mathcal{C}$ be the set of atomic concepts (or labels) to learn. This paper main idea comes from the fact that the concepts to be learned are not totally independent, thus grouping some concepts to learn them against others using implicit biases considerably improves the quality of learning for each concept. For instance, learning *take the subway* and *take the train* together against all other concepts results better than learning each of these two concepts separately. This is because the bias separating these two concepts from all the others is not the same as that which separates them from each other. This strong and effective idea is the basis of this work. We define a space of concept groups where new concepts are defined from similar ones and learned using specialized classifiers. If any concept group improves the learning rate of related atomic concepts, then it will appear in the final learning structure. The main problem will be to find the best structure of directed concepts groups to be learned in order to optimize the learning rate of each atomic concept.

Note that in this setting, according to the three dimensions defined in [12], we consider: (1) single-label classification as opposed to multi-label classification; (2) the type of hierarchy (or structure) to be trees as opposed to directed acyclic graphs; (3) instances that have to be classified into leafs, i.e. mandatory leaf node prediction [20], as opposed to the setting where instances can be classified into any node of the hierarchy (early stopping).

A tree hierarchy organizes the class labels into a tree-like structure to represent a kind of "IS-A" relationship between labels. Specifically, [12] points out that the properties of the "IS-A" relationship can be described as asymmetry, anti-reflexivity and transitivity [20]. We define a tree as a pair $(\mathcal{C}, \prec)$, where $\mathcal{C}$ is the set of class labels and "$\prec$" denotes the "IS-A" relationship.

Let $\{(x_1, c_1), \ldots, (x_N, c_N)\} \overset{\text{i.i.d.}}{\sim} X, C$ be a set of training examples, where X and C are two random variables taking values in $\mathcal{X} \times \mathcal{C}$, respectively. Each $x_k \in \mathcal{X}$ and each $c_k \in \mathcal{C}$. Our goal is to learn a classification function $f : \mathcal{X} \to \mathcal{C}$ that attains a small classification error. In this paper, we associate each node i with a classifier $\mathcal{M}_i$, and focus on classifiers $f(x)$ that are parameterized by $\mathcal{M}_1, \ldots, \mathcal{M}_m$ through the following recursive procedure [34]:

$$f(x) = \begin{cases} \textbf{initialize } i := 0 \\ \textbf{while } \mathcal{C}(i) \text{ is not empty} \\ \qquad i := \operatorname{argmax}_{j \in Child(i)} \mathcal{M}_j(x) \\ \textbf{return } i \end{cases} \tag{1}$$

where $Child(i)$ is the set of children for the node i.

In the case of the SHL dataset, for instance, learning *train* and *subway* or *car* and *bus* before learning each one alone gives better results. Advantages of considering these classes paired together as opposed to the flat classification setting which leads to significant degradation of recognition performances as demonstrated in some works around the SHL dataset [27]. In contrast, organizing the various concepts into a tree-like

[1] In our case, among the 16 input modalities of the original dataset, we select the body-motion modalities to be included in our experiments: *accelerometer*, *gyroscope*, etc. Additional details on the segmentation and processing of the inputs can be found in the experimental part.

structure, inspired by domain expertise, demonstrated significant gains in terms of recognition performances in the context of the SHL challenge [14] and activity recognition in general [18, 19].

Designing such structures is of utmost importance but hard because it involves optimizing the structure as well as the weights of the classifiers attached to the nodes of that structure (See Section 2.2). Our goal is then to determine an optimal structure of classes that can facilitate (improve and accelerate) learning of the whole concepts. After finding the most fitted structure, the next challenge is to find the best combination in order to learn the whole concepts with a minimum value of error.

2.2 Search Space Size: Complexity Analysis

A naive approach is to generate the lattice structure of concepts groups and to choose the tree hierarchies which give the best accuracy of atomic concepts. In practice, this is not doable because of the exponential (in the number of leaf nodes) number of possible trees. We propose a recurrence relation involving binomial coefficients for calculating the total number of tree hierarchies for K different concepts (class labels).

Example 1 *As an example assume we have* 3 *various concepts, we are interested in counting the total number of hierarchies for classifying these concepts. We consider that we have three classes namely* c_1, c_2 *and* c_3*, there exist* 4 *different tree hierarchies for learning the classification problem as below:*
– $(c_1c_2c_3)$*: the tree has one level and the learning process takes one step. Three concepts are learned while each concept is learned separately from the others (flat classification).*
– $((c_1c_2)c_3)$*: the tree has two levels and the learning process takes two steps: at the first level, it learns two concepts (atomic* c_3 *and two atomics* c_2 *and* c_3 *together) . At the second level it learns separately the two joined concepts* c_2 *and* c_3 *of the first level, etc.*
– $(c_1(c_2\ c_3))$ *and* $((c_1c_3)c_2)$*.*

Theorem 1 *Let* $L(K)$ *be the total number of trees for the given* K *number of concepts. The total number of trees for* $K+1$ *concepts satisfies the following recurrence relation:* $L(K+1) = \binom{K}{K-1}L(n)L(1) + 2\sum_{i=0}^{K-2}\binom{K}{i}L(i+1)L(K-i)$.

Proof: It can be explained by observing that, for $K+1$ concepts containing K existed concepts $c_1, \cdots c_K$ and a new added concept γ, we can produce the first level trees combinations as below. Notice that each atomic element o can be one of the $c_1, \cdots c_K$ concepts. In order to compute the total number of trees combinations, we show what is the number of tree combinations by assigning the K concepts to each item:

- $(\gamma(\overbrace{o\cdots o}^{K\text{concepts}}))$: the number of trees combinations by taking the concept labels into the account are: $\binom{K}{0}L(1)\times 2\times L(K)$; the reason for multiplying the number of trees combinations for K concepts to 2 is because while the left side contains an atomic γ concept, there are two choices for the right side of the tree in the first level: either we compute the total number of trees for K concepts from the first level or we keep the first level as a $\overbrace{o\cdots o}^{K\text{concepts}}$ atomics and keep all K concepts together, then continue the number of K trees combinations from the second level of the tree.
- $((\gamma o)(\overbrace{o\cdots o}^{K-1\text{concepts}}))$: similar to the previous part we have $\binom{K}{1}L(2)\times 2\times L(K-1)$ trees combinations by taking the concepts labels into the account. $\binom{K}{1}$ indicates the number of combinations for choosing a concept from the K concept and put it with the new concept separately. While $L(2)$ is the number of trees combinations for the left side of tree separated with the new concept γ.
- $((\gamma oo)(\overbrace{o\cdots o}^{K-2\text{concepts}}))$, $\cdots$
- $((\gamma\ \overbrace{o\cdots o}^{K-1\text{concepts}}\ o)$: $\binom{K}{K-1}L(n)L(1)$ in this special part, we follow the same formula except the single concept in the right side has only one possible combination in the first level equal to $L(1)$.

All in all, the sum of these items calculates the total number of tree hierarchies for $K+1$ concepts. □
The first few number of total number of trees combinations for $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, \cdots$ concepts are: 1 , 1, 4, 26, 236, 2752, 39208, 660032, 12818912, 282137824, $\cdots$. In the case of the SHL dataset that we use in the empirical evaluation, we have 8 different concepts and thus, the number of different types of hierarchies for this case is $L(8) = 660,032$.

3 Proposed Approach

Our goals are to: (i) organize the considered concepts into hierarchies such that the learning process accounts for the dependencies that exist among these concepts; (ii) characterize optimal classifiers (Empirical Risk Minimizers (ERMs)) [25] that are associated to each non-leaf node of the hierarchies. Structuring the concepts can be performed using two different approaches: a **top-down** approach where we seek to decompose the learning process; a **bottom-up** approach

where specialized models are grouped together based on their affinities. Our approach takes the latter direction and constructs hierarchies based on the similarity between concepts. We keep adapting the constructed hierarchies as we continuously analyze the similarity by increasing the transfer orders and the supervision budget. At each time, the classifiers assigned to the nodes of the constructed hierarchy are refined recursively in order to take into account the hierarchical dependence (global loss). In this section, we detail the different parts of our approach which are illustrated in Figure 1. Next we will introduce the three stages of our approach: *Concept similarity analysis*, *Hierarchy derivation*, and *Hierarchy refinement*.

3.1 Concept Affinity Analysis

In our **bottom-up** approach we leverage transferability and dependency among concepts as a measure of similarity. Besides the nice empirical properties of this measure (see **Properties**), the rationale behind it has to do with the re-usability of what was learned so far in lower levels of the hierarchies. Indeed, we leverage the models that we learned during this step and use them with few additional adjustments in the final hierarchical learning setting.

Transfer-based affinity. Given the set of concepts $\mathcal{C}$, we compute during this step an affinity matrix that captures the notion of transferability, and by the same occasion that of similarity, among concepts (See **Appropriateness of the transfer-based affinity measure**). For this, we first compute for each concept $c_i \in \mathcal{C}$ an encoder $f_\theta^{c_i}$ (parameterized by θ) that learns to map inputs, labeled as c_i, to representation $\mathcal{Z}_{c_i}$. Learning the parameters of the encoder consists in minimizing the reconstruction error, that is, carrying the following optimization [26] $\text{argmin}_{\theta,\theta'} \mathbb{E}_{x,c\sim X,C|c=c_i} \mathcal{L}(g_{\theta'}^{c_i}(f_\theta^{c_i}(x)), x)$, where $g_{\theta'}^{c_i}$ is a decoder which maps back the learned representation into the original inputs space.

We propose to leverage the learned encoder, for a given concept c_i, to compute affinities with other concepts via fine-tuning of the learned representation. Precisely, we fine-tune the encoder $f_{c_i}^\theta$ to account for a target concept $c_j \in \mathcal{C}$. This process consists, similarly, in minimizing the reconstruction error however rather than using the decoder $g_{\theta'}^{c_i}$ learned above we design a genuine decoder $g_{\theta'}^{c_j}$ that we learn from scratch. The corresponding objective function is $\text{argmin}_{\theta,\theta'} \mathbb{E}_{x,c\sim X,C|c=c_j} \mathcal{L}(g_{\theta'}^{c_j}(f_\theta^{c_i}(x)), x)$. We use the performance of this step as a *similarity score* from c_i to c_j which we denote by $p_{c_i \to c_j} \in [0, 1]$. We refer to the number of examples belonging to the concept c_j used during fine-tuning as the *supervision budget*, denoted b, which is used to index a given measure of similarity. It allows us to have an additional indicator as to the similarity between the considered concepts. The final similarity score is computed as $\frac{\alpha \cdot p_{c_i \to c_j} + \beta \cdot b}{\alpha + \beta}$. We set α and β to be equal to $\frac{1}{2}$.

Properties. In many applications, e.g. computer-vision [31] and natural language processing [16], variants of the transfer-based affinity have been shown empirically to improve (i) the **quality** of transferred models (wins against fully supervised models) as well as (ii) **gains**, i.e. win rate against a network trained from scratch using the same training data as transfer networks'. And more importantly (iii) **universality** of the resulting structure. Indeed, the affinities based on transferability are stable despite the variations of a big corpus of hyperparameters. In addition, we provide empirical evidence (See Sect. 4.2) of the appropriateness of the transfer-based affinity measure as a for the separability of the similar concepts and on the contrary the difficulty to separate concepts that exhibit low similarity score.

3.2 Hierarchy Derivation

Given the set of *affinity scores* obtained previously, we derive the most appropriate hierarchy, following an agglomerative clustering method combined with some additional constraints.

The agglomerative clustering method proceeds by a series of successive fusions of the concepts into groups and results in a structure represented by a two-dimensional diagram known as a dendogram. It works by (1) forming groups of concepts that are close enough and (2) update the affinity scores based on the newly formed groups. This process is defined by the recurrence formula proposed by [13] and gives the distance between a group (of concepts) k and a group (ij) formed by the fusion of two groups (i and j) as $d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}|$, where d_{ij} is the distance between groups i and j. By varying the parameter values $\alpha_i, \alpha_j, \beta$, and γ, we expect to get clustering schemes with various characteristics.

In addition to the above updating process, we propose additional constraints to refine further the hierarchy derivation stage. Given the dendogram produced by the agglomerative method above, we define an *affinity threshold* τ such that if the distance at a given node is $d_{ij} \geq \tau$, then we merge the nodes to form a unique subtree. In addition, as we keep track of the quantities of data used to train and fine-tune the encoders during the transfer-based affinity analysis stage, this indicator is exploited to inform us as to which nodes to merge.

Let $\mathcal{T}$ be the derived hierarchy (tree) and let t index

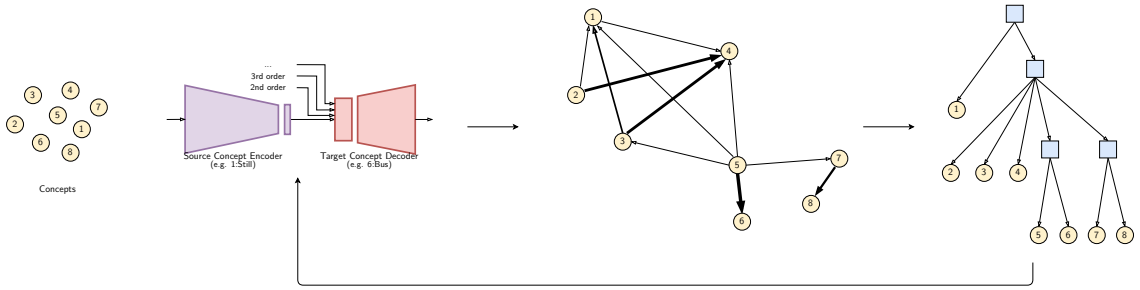

Figure 1: The framework of the proposed approach involving 3 major steps: (1) Concept similarity analysis: encoders are trained to output, for each source concept, an appropriate representation which is then fine-tuned in order to account for target concepts. Affinity scores are depicted by the arrows between concepts (the thicker the arrow, the higher the affinity score). (2) Hierarchy derivation: based on the obtained affinity scores, a hierarchy is derived using an agglomerative approach. (3) Hierarchy refinement: each non-leaf node of the derived hierarchy is assigned with a model that encompasses the most appropriate representation as well as an ERM which is optimized to separate the considered concepts. The process is repeated until convergence.

the non-leaf or internal nodes. The leafs of the hierarchy correspond to the considered concepts. For any non-leaf node t, we associate a model $\mathcal{M}_t$ that encompasses (1) an encoder (denoted in the following simply by $\mathcal{Z}_t$ in order to focus on the representation) that maps inputs X to representations $\mathcal{Z}_t$ and (2) an ERM f_t (such as support vector machines SVMs) that outputs decision boundaries based on the representations produced by the encoder.

3.3 Hierarchy Refinement

After explaining the hierarchy derivation process, in this part, we will discuss: (1) which representations are used in each individual model; and (2) how each individual model (including the representation and the ERM weights) is adjusted to account for both the local errors and also those of the hierarchy as a whole.

Which representations to use? The question discussed here is related to the encoders to be used in each non-leaf node. For any non-leaf node t we distinguish two cases: (i) all its children are leafs; (ii) it has at least one non-leaf node. In the first case, the final representation that is considered by the ERM associated with the non-leaf node is the representation learned in the concept affinity analysis step (first-order transfer-based affinity). In the second case, we can either fuse the nodes (For example, we get a classification between 3 concepts together rather than {1} vs. {2,3} then {2} vs. {3}. See Figure 2) or keep it as it is and leverage the affinities based on higher-order transfer where, rather than accounting for a unique target concept, the representation is fine-tuned to account for each one of the concepts involved in the considered node. Figure 2 illustrates examples of how transfers are performed between the models associated with the non-leaf nodes. We index the models with the encoder that compose it like this $\mathcal{M}_{[\mathcal{Z}_i]}$. In the case of higher-order transfer, the models are indexed using all concepts involved in the transfer, i.e. $\mathcal{M}_{[\mathcal{Z}_{i,j,\ldots}]}$.

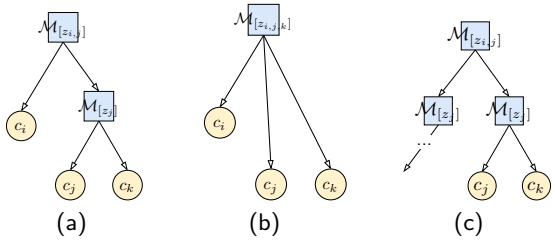

Figure 2: Examples of how transfers are performed between the models associated with the non-leaf nodes. the models of the hierarchy in (a) can be kept as they are merged to form the hierarchy in (b). In the second case, we need to perform a high-order transfer between the concepts c_i, c_j, and c_k. (c) Situation where no transfers can be made.

Adjusting weights of the models. These classifiers are trained to output decision boundaries (a hypothesis) based on the most appropriate representations learned earlier. Given the encoder (representation) assigned to any non-leaf node t, we select a classifier $\hat{f} := \text{argmin}_{f \in \mathcal{H}} \hat{R}(f, \mathcal{Z}_t)$ where $\hat{R}(f, \mathcal{Z}_t) := \frac{1}{M} \sum_{x,c \sim X,C|c \in Child(t)} \mathbb{E}_{z \sim \mathcal{Z}_t|x}[\mathcal{L}(c, f(z))]$ and $\mathcal{H}$ is the hypothesis space. Models are adjusted to account for local errors as well as for global errors related to the hierarchy as a whole. In the first case, the loss is defined as the traditional hinge loss used in SVMs which is intended to adjust the weights of the classifiers that have leafs as direct children. In the second case, we use a loss that encourages the models to leverage orthogonal representations (between children and parent nodes) [34].

4 Experiments and Results

In this section, we perform an empirical evaluation of the proposed approach and consisting of three steps: In the first stage, we evaluate classification performances in the hierarchical setting (Sect. 4.1); we, then, evaluate the transfer-based affinity analysis step and the properties related to the separability of the considered concepts (Sect. 4.2); Finally, we evaluate the derived hierarchies in terms of stability, performance, and agreement with their counterparts defined by domain experts. (Sect. 4.3) The code to reproduce the experiments is publicly available [2].

SHL dataset. The SHL dataset [8] is a highly versatile and precisely annotated dataset dedicated to mobility-related human activity recognition. In contrast to related representative datasets like [33, 32, 30, 3], the SHL dataset provides, simultaneously, multimodal and multilocation locomotion data recorded in real-life settings. There are in total 16 modalities including accelerometer, gyroscope, cellular networks, WiFi networks, audio, etc. making it suitable for a wide range of applications. Data collection was performed by each participant using four smartphones simultaneously placed in different body locations: *Hand*, *Torso*, *Hips*, and *Bag*. These four positions define the topology that allows us to model and leverage the dynamics of body movements for activity recognition models. Among the 16 modalities of the original dataset, we select the body-motion modalities to be included in our experiments, namely: *accelerometer*, *gyroscope*, *magnetometer*, *linear acceleration*, *orientation*, *gravity*, and in addition, *ambient pressure*. This makes the data set suitable for a wide range of applications and in particular the task we are interested in, in this paper which is transportation recognition. Indeed, there are 8 primary categories of transportation that we are interested in: *1:Still*, *2:Walk*, *3:Run*, *4:Bike*, *5:Car*, *6:Bus*, *7:Train*, and *8:Subway (Tube)*.

Training details. We use Tensorflow [1] for building the encoders (and decoders). In this work, we construct encoders by stacking Conv1d/ReLU/MaxPool blocks. These blocks are followed by a Fully Connected/ReLU layers. Encoders performance estimation is based on the validation loss and is framed as a sequence classification problem. As a preprocessing step, annotated input streams from the SHL dataset are segmented into sequences of 6000 samples which correspond to a duration of 1 min. given a sampling rate of 100 Hz. For weight optimization, we use stochastic gradient descent with Nesterov momentum of 0.9 and a learning-rate of 0.1 for a minimum of 12 epochs (we stop training if there is not improvement). Weight decay is set to 0.0001. Furthermore, to make the neural networks more stable, we use batch normalization on top of each convolutional layer [11]. We use SVMs as our ERMs in the derived hierarchies.

Evaluation Metrics. In hierarchical classification settings, the hierarchical structure is of utmost importance and should be taken into account during model evaluation [20]. Various measures that account for the hierarchical structure of the learning process have been studied in the literature. These can be categorized into: distance-based; depth-dependent; semantics-based; hierarchy-based measures, each displaying advantages and disadvantages depending on the characteristics of the considered structure [6]. In our experiments, we use the *H-loss*, a hierarchy-based measure defined in [4]. This measure captures the intuition that *"whenever a classification mistake is made on a node of the taxonomy, then no loss should be charged for any additional mistake occurring in the subtree of that node."* $\ell_H(\hat{y}, y) = \sum_{i=1}^{N} \{\hat{y} \neq y \wedge \hat{y}_j = y_j, j \in Anc(i)\}$, where $\hat{y}$ is the predicted and $Anc(i)$ is the set of ancestors of the node i.

4.1 Evaluation of the Hierarchical Classification Performances

In this set of experiments, we evaluate the flat classification setting which constitutes our baseline for the rest of the empirical evaluations. To make our baseline comparable with the hierarchical models, we make sure to get the same complexity, i.e. comparable number of parameters as the largest hierarchies including the weights of the encoders and those of the ERMs. We use Bayesian optimization based on Gaussian processes as surrogate models to select the optimal hyperparameters of the baseline model [21]. Additional details about the baseline model and the hyperparameters used to obtain the baseline model can be found in the code repository.

Per-node performances. Figure 3 shows the resulting per-node performances, i.e. how accurately the models associated with the non-leaf nodes can predict the correct subcategory? averaged over the entire derived hierarchies. The nodes are ranked according to the obtained per-node performance (top 10 nodes are shown) and accompanied by their appearance frequency. It is worth noticing that the concept 1:*still* learned alone against the rest of the concepts (first bar) achieves the highest gains in terms of recognition performances while the appearance frequency of this learning configuration is high (more than 60 times).

[2] Software package and code to reproduce empirical results are publicly available at https://github.com/alphaequivalence/hierarchicalSHL

We see also that the concepts 4:*bike*, 5:*car*, and 6:*bus* grouped together (5th bar) occur very often in the derived hierarchies (80 times) which is accompanied by fairly significant performance gains ($5.09 \pm 0.3\%$). At the same time, as expected, we see that the appearance frequency gets into a plateau starting from the 6th bar (which lasts after the 10th bar). This suggests that the most influential nodes are often exhibited by our approach.

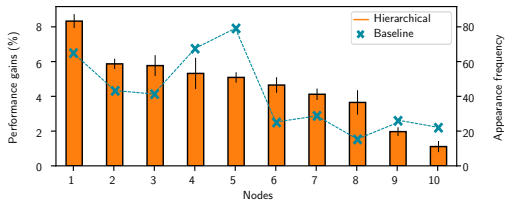

Figure 3: Per-node performance gains averaged over the entire derived architectures (similar nodes are grouped and their performances are averaged). The appearance frequency of the nodes is also illustrated. For each bar, the corresponding concepts can be found in the code repository. For example, the 8th bar corresponds to the concepts 2:*walk*-3:*run*-4:*bike* grouped.

Per-concept performances. We further ensure that the performance improvements we get at the node levels are reflected at the concept level. Figure 4 shows the recognition performances of each concept averaged over the whole hierarchies derived using our proposed approach. We indeed observe that there are significant improvements for each individual concept over the baseline (flat classification setting). We can observe that again 1:*still* has the highest classification rate ($72.32 \pm 3.45\%$) and an improvement of 5 points over the baseline. Concept 6:*bus* also exhibits a roughly similar trend. On the other hand, concept 7:*train* has the least gains ($64.43 \pm 4.45\%$) with no significant improvement over the baseline. Concept 8:*subway* exhibits the same behavior suggesting that there are undesirable effects that stem from the definition of these two concepts.

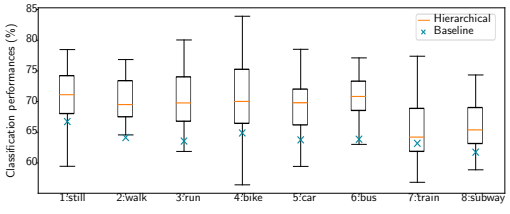

Figure 4: Recognition performances of each individual concept averaged over the entire derived hierarchies. For reference, the recognition performances of the baseline model are also shown.

4.2 Evaluation of the Affinity Analysis Stage

In these experiments, we evaluate the proposed transfer-based affinity measure. We assess, the separability of the concepts depending on their similarity score (which accounts for both the transfer-affinity and supervision budget) and the learned representation.

Appropriateness of the Transfer-based Affinity Measure. We reviewed above the nice properties of the transfer-based measure especially the universality and stability of the resulting affinity structure. The question that arises is related to the separability of the concepts that are grouped together. Are the obtained representations optimal for the final ERMs that are used for classification? This is what we investigate here. Figure 5 shows the decision boundaries generated by the considered ERMs which are provided with the learned representations of two concepts. The first case (top right), exhibits a low-affinity score, and the second case (bottom right) shows a high-affinity score. We can see, in this example, that the boundaries in the first case are unable to separate the two concepts while it gets a fairly distinct frontier.

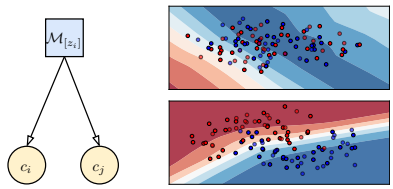

Figure 5: Decision boundaries generated by the considered ERMs using two different encoders (representations) fine-tuned to account for (top right) a dissimilar concept (exhibiting a low-affinity score); (bottom right) a similar concept (exhibiting a high-affinity score).

Impact on the ERMs' Decision Boundaries. We train different models with various learned representations in order to investigate the effect of the initial affinities (obtained solely with a set of 100 learning examples) and the supervision budget (additional learning examples used to fine-tune the obtained representation) on the classification performances of the ERMs associated with the non-leaf nodes of our hierarchies. Figure 6 shows the decision boundaries generated by various models as a function of the distance between the concepts (y-axis) and the supervision budget (x-axis). Increasing the supervision budget to some larger extents (more than ~ 300 examples) results in a substantial decrease in classification performances of the ERMs. This suggests that, although our initial affinity scores are decisive (e.g. 0.8), the supervision

budget is tightly linked to generalization. This shows that a trade-off (controlled by the supervision budget) between separability and initial affinities arises when we seek to group concepts together. In other words, this arises the question of whether to increase the supervision budget indefinitely (in the limits of available learning examples) in order to find the most appropriate concepts to fuse with, while expecting good separability.

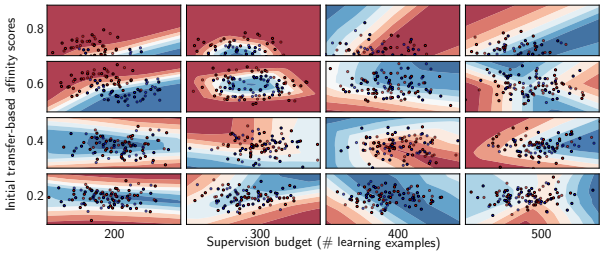

Figure 6: Decision boundaries obtained by SVM-based classifiers trained on the representations $\mathcal{Z}_t$ as a function of the distance between the concepts (y-axis) and the supervision budget (x-axis).

4.3 Universality and Stability

We demonstrated in the previous section the appropriateness of the transfer-based affinity measure to provide distance between concepts as well as the existence of a trade-off between concepts separability and their initial affinities. Here we evaluate the **universality** of the derived hierarchies as well as their **stability** during adaptation with respect to our hyperparameters (affinity threshold and supervision budget).

We compare the derived hierarchies with their domain experts-defined counterparts as well as those obtained via a random sampling process. Figure 7 shows some of the hierarchies defined by the domain experts (first row) and sampled using the random sampling process. For example, the hierarchy depicted in Figure 7d corresponds to a split between static (1:*still*, 5:*car*, 6:*bus*, 7:*train*, 8:*subway*) and dynamic (2:*walk*, 3:*run*, 4:*bike*) activities. The difference between the hierarchies depicted in Figure 7a and 7b is related to 4:*bike* activity which is linked first to 2:*walk* and 3:*run* then to 5:*car* and 6:*bus*. A possible interpretation is that in the first case, biking is considered as "on feet" activity while in the second case as "on wheels" activity. What we observed is that the derived hierarchies tend to converge towards the expert-defined ones.

We compare the derived hierarchies in terms of their level of agreement. We use for this assessment, the Cohen's kappa coefficient [5] which measures the agreement between two raters. The first column of Table 1

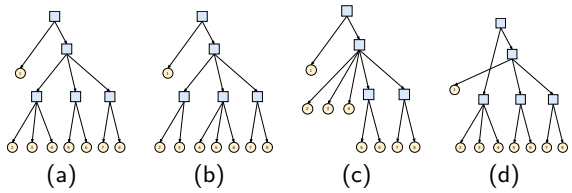

Figure 7: Hierarchies: (a) defined by domain expert (b-c) derived using our approach (d) randomly sampled.

provides the obtained coefficients. We also compare the average recognition performance of the derived hierarchies (second column of Table 1).

Table 1: Summary of the recognition performances obtained with our proposed approach compared to randomly sampled and expert-defined hierarchies. Comparison performed in terms of agreement (agr.) and average performance (perf.).

Method	Agr.	perf. avg.± std.
Expertise	-	72.32±0.17
Random	0.32	48.17±5.76
Proposed	0.77	75.92±1.13

In terms of stability, as we vary the design choices (hyperparameters), defined in our approach, we found that the affinity threshold has a substantial impact on our results with many adjustments involved (12 hierarchy adjustments on avg.) whereas the supervision budget has a slight effect, which confirms the observations in Sect. 4.2.

5 Conclusion and Future Work

In this work, we propose an approach for organizing the learning process of dependent concepts in the case of human activity recognition. We first determine a suitable structure for the concepts according to a transfer affinity-based measure. We then characterize optimal representations and classifiers which are then refined to account for both local and global errors. We provide theoretical bounds for the problem and empirically show that using our approach we are able to improve the performances and robustness of activity recognition models over a flat classification baseline. In addition to supporting the necessity of organizing concepts learning, our experiments raise interesting questions for future work. Noticeably, results in Sect. 4.2 asks what is the optimal amount of supervision for deriving the hierarchies. Another future work is to study different approaches for searching and exploring the search space of different hierarchical types (trees, lattices, etc.) to extract new classification methods and to compare our results with them.

References

[1] M. Abadi et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 78–87, 2004.

[3] C. Carpineti, V. Lomonaco, L. Bedogni, M. Di Felice, and L. Bononi. Custom dual transportation mode detection by smartphone devices exploiting sensor diversity. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 367–372. IEEE, 2018.

[4] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Incremental algorithms for hierarchical classification. *Journal of Machine Learning Research*, 7(Jan):31–54, 2006.

[5] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] E. Costa, A. Lorena, A. Carvalho, and A. Freitas. A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods for machine Learning II: papers from the AAAI-2007 Workshop*, pages 1–6, 2007.

[7] M. Essaidi, A. Osmani, and C. Rouveirol. Learning dependent-concepts in ilp: Application to model-driven data warehouses. In *Latest Advances In Inductive Logic Programming*, pages 151–172. World Scientific, 2015.

[8] H. Gjoreski, M. Ciliberto, L. Wang, F. J. Ordonez Morales, S. Mekki, S. Valentin, and D. Roggen. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. *IEEE Access*, 2018.

[9] H. Gjoreski and D. Roggen. Unsupervised online activity discovery using temporal behaviour assumption. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 42–49, 2017.

[10] M. Hamidi and A. Osmani. Data generation process modeling for activity recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37, pages 448–456. PMLR, 2015.

[12] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

[13] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies: 1. hierarchical systems. *The computer journal*, 9(4):373–380, 1967.

[14] Y. Nakamura, Y. Umetsu, J. P. Talusan, K. Yasumoto, W. Sasaki, M. Takata, and Y. Arakawa. Multi-stage activity inference for locomotion and transportation analytics of mobile users. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1579–1588, 2018.

[15] L.-V. Nguyen-Dinh, A. Calatroni, and G. Tröster. Robust online gesture recognition with crowdsourced annotations. *The Journal of Machine Learning Research*, 15(1):3187–3220, 2014.

[16] M. E. Peters, S. Ruder, and N. A. Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*, 2019.

[17] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, et al. Dataset issues in object recognition. In *Toward category-level object recognition*, pages 29–48. Springer, 2006.

[18] F. Samie, L. Bauer, and J. Henkel. Hierarchical classification for constrained iot devices: A case study on human activity recognition. *IEEE Internet of Things Journal*, 2020.

[19] S. Scheurer, S. Tedesco, K. N. Brown, and B. O'Flynn. Using domain knowledge for interpretable and competitive multi-class human activity recognition. *Sensors*, 20(4):1208, 2020.

[20] C. N. Silla and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

[21] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *NIPS*, pages 2951–2959, 2012.

[22] M. Stikic and B. Schiele. Activity recognition from sparsely labeled data using multi-instance learning. In *International Symposium on Location-and Context-Awareness*, pages 156–173. Springer, 2009.

[23] V. Taran, Y. Gordienko, A. Rokovyi, O. Alienin, and S. Stirenko. Impact of ground truth annotation quality on performance of semantic image segmentation of traffic conditions. In *International Conference on Computer Science, Engineering and Education Applications*, pages 183–193. Springer, 2019.

[24] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009.

[25] V. Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

[26] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[27] L. Wang, H. Gjoreskia, K. Murao, T. Okita, and D. Roggen. Summary of the sussex-huawei locomotion-transportation recognition challenge. In *Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers*, pages 1521–1530, 2018.

[28] J. Wehrmann, R. Cerri, and R. Barros. Hierarchical multi-label classification networks. In *International Conference on Machine Learning*, pages 5075–5084, 2018.

[29] H. Yao, Y. Wei, J. Huang, and Z. Li. Hierarchically structured meta-learning. In *International Conference on Machine Learning*, pages 7045–7054, 2019.

[30] M.-C. Yu, T. Yu, S.-C. Wang, C.-J. Lin, and E. Y. Chang. Big data small footprint: the design of a low-power classifier for detecting transportation modes. *Proceedings of the VLDB Endowment*, 7(13):1429–1440, 2014.

[31] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

[32] M. Zhang and A. A. Sawchuk. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 1036–1043, 2012.

[33] Y. Zheng, X. Xie, and W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[34] D. Zhou, L. Xiao, and M. Wu. Hierarchical classification via orthogonal transfer. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 801–808, 2011.