ILPC: Iterative Learning using Physical Constraints in Real-World Sensing Data

Tong Yu[1,*], Shijia Pan[1,*], Susu Xu[2], Xinlei Chen[1], Mostafa Mirshekari[2], Jonathon Fagert[2], Hae Young Noh[2], Pei Zhang[1], and Ole J. Mengshoel[1]

1. Dept. of Electrical and Computer Eng., Carnegie Mellon Univ., Moffett Field, CA, USA, 94035
2. Dept. of Civil and Environmental Eng., Carnegie Mellon Univ., Pittsburgh, PA, USA, 15213
* equal contribution among authors.

Abstract

In this paper, we present Iterative Learning using Physical Constraints (ILPC) method. ILPC is an iterative learning method targeting at model inaccuracy caused by a distribution change in training and test data. This change in distribution can be due to the complexity of many real-world physical systems. Although domain adaptation methods, which consider both training and test data distribution when building models, also target this distribution change, these methods can only handle a limited difference between training and test data.

ILPC handles different distributions based on a key observation: gradual changes in physical condition often cause gradual data distribution changes. Instead of treating test data as generated by an identical distribution, ILPC builds a model iteratively, guided by a system's physical measurements. In each iteration, the model is only extended with data that has similar physical measurements to the last iteration. This approach leads to higher accuracy. To evaluate ILPC, we apply it to two real-world datasets and achieve up to a 2.7× improvement in prediction accuracy compared to existing domain adaptation methods.

Introduction

With the growth of the smart devices, cyber-physical systems (CPS), and Internet of Things systems (IoT) become ubiquitous in everyday lives. Many examples of these IoT devices exist (e.g., Samsung Smartthings, Notions (Samsung Inc. 2017; Notion Inc. 2017)) and their number is expected to reach 24 billion by the year 2020 (Gubbi et al. 2013). These systems with their varying sensing abilities can extract information of both physical environments and humans in the environments for various IoT applications, including smart buildings. However, machine learning models are often inaccurate when applied to these real world-deployed sensing systems. A common cause of these inaccurate models is the distribution change between the training and test data due to the complexity of the physical world (Shi and Sha 2012; Pan et al. 2011).

Previous approaches to build more generalizable models for changing distributions are transductive learning and domain adaptation (Chapelle, Schölkopf, and Zien 2010; Joachims 1999b; Zhu et al. 2003; Shi and Sha 2012). These methods consider the distribution of unlabeled test data when training a model. They train the model in one shot, and they operate under the assumption that there is a great similarity between training and test distributions, i.e., the difference between training and test distributions is not significant. However, in real-world sensing systems, complex physical conditions may cause training and test data to have very different distributions. As a result, existing methods are not robust to test data with a significantly different distribution than training data. We make these key observations on the data distribution changes in physical sensing systems: 1) When certain measurable physical constraints change, the data distribution often also changes. 2) If these physical constraints change gradually, the corresponding data distribution typically also changes gradually.

Based on these key observations, we present Iterative Learning using Physical Constraints (ILPC), an iterative method for learning guided by physical measurements. ILPC initializes with labeled data of limited distribution range. Then in each iteration, ILPC uses an existing domain adaptation on test data with similar distributions to the labeled data, which is selected based on the measurable physical constraints, to ensure high prediction accuracy. The iterative results with high prediction confidence are then 'labeled', so that the approximately labeled data distribution expands. In the next iteration, ILPC can handle a larger difference in data distribution than that of the current iteration. Compared to traditional methods of transductive learning and domain adaptation, our ILPC handles significant data distribution changes by iteratively extending the model. It ensures the iterative accuracy by taking corresponding physical measurement into account when selecting data trained in an iteration. The contributions of this paper are as follows.

- We study the physical constraints leading to gradual data distribution changes in various IoT sensing applications.
- We present a novel iterative approach, Iterative Learning using Physical Constraints (ILPC), to handle significant distribution change problems caused by gradually changing physical constraints.
- We apply ILPC to two real-world datasets and achieve more than 30% test accuracy improvements compared to existing domain adaptation methods.

Gradual Data Distribution Changes from Physical Constraints

The data distribution changes in many real-world systems are caused by physical condition changes, and these conditions

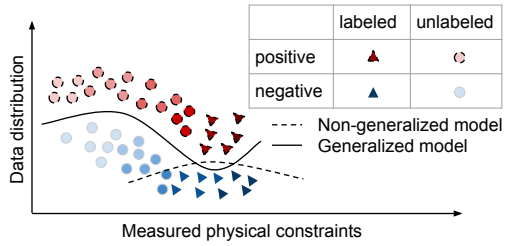

Figure 1: Examples of changes in data distribution. The x-axis is the measured physical constraint. The y-axis is the data distribution as it gradually changes with the measured physical constraint. The model represented by the dashed line does not generalize while the model represented by the solid line separates the data with a large difference in distribution, which is the goal of this work.

can be measured. We refer to these measurable physical conditions as *physical constraints*. In this section, we study the above cause-effect relationship in two example datasets and explain why it is difficult for prior work to handle significant data distribution changes.

Gradual Distribution Changes in Physical Systems

In real-world cyber-physical systems, the training and test sensing data samples may be collected under different physical conditions. When that happens, the underlying distributions generating the data samples are different. Figure 1 shows a concept example where the distributions are different for data collected under different physical attribute values. The blue triangles and circles are negative data samples, while the red triangles and circles are positive data samples. The x-axis shows the values of a measured physical constraint, and the y-axis shows the data distribution under the physical constraint values. Although there are distribution changes between different data samples, these changes appear to be gradual and are linked to the changes in the physical constraint values. Furthermore, in physical sensing systems, we can identify and measure these physical constraint values. This phenomenon can be found in real-world sensing data. We introduce two applications that demonstrate different physical constraints.

Applications and Their Physical Constraints

Pedestrian identification through floor vibration Occupant identity is useful information for various smart building applications, and one way to obtain this is through footstep-induced floor vibration sensing. The frequency domain of the vibration signal is extracted for identification using classifiers such as SVM (Pan et al. 2015). However, the identification accuracy drops when people walk at speeds different from that in labeled dataset. This is because when their walking speeds change, their gaits and therefore their footstep signal features also change. Figure 2 shows an example of changes in one person's footstep signal distribution over several different walking speeds. The x-axis represents 7 walking speeds from slow to fast. All the footsteps are clustered based on their time domain similarity into three types. The y-axis shows

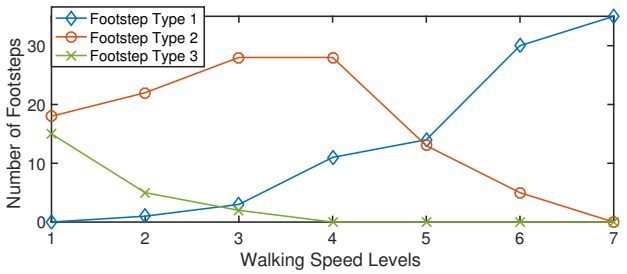

Figure 2: Footstep example. The figure shows footstep signals collected from one person walking at seven different speeds (1-7 indicates speeds from slow to fast). Each speed has 35 footstep samples. The footsteps are clustered based on their time domain similarity into three types. Footstep Type 1, 2, and 3 appear more often in the fast, medium, and slow walking speeds respectively. As the walking speed changes, the frequency footstep type also gradually changes.

the number of footsteps that fall into each type. When the walking speed increases from 1 to 7, the number of Type 1 footsteps increases gradually, while the number of Type 3 steps decreases gradually. The Footstep Type 2 peaks at walking speed 3 and decreases when the walking speed is above 4. As the walking speed changes, the distribution of the footstep data also gradually changes.

Building damage estimation via mid-earthquake structural vibrations Evaluating the structural damage level of a building post-earthquake is essential in helping to save people and reduce economic loss. To diagnose building damage, people has been studied the building structural vibration during the earthquake (Xu, Zhang, and Noh 2017). Features in the frequency domain of the vibration signals can be extracted and used to train classification models. With those models, we can predict the state of each story and localize structural damage inside a building. However, the estimation accuracy decreases when earthquake intensities varies between training and test data. This is because the degree of building damage changes at different intensity levels, resulting in different data distributions. Therefore, a key physical constraint in this case is the earthquake intensity level.

The Problem of Distribution Changes

While the pedestrian identification and building damage estimation application differ in many ways, in both cases the distribution of testing data changes based on a specific and measurable physical constraint. We can use this observation to further analyze the learning problem with prior attempts of domain adaptation to compensate for data distribution changes. In supervised machine learning, if training and test data are from different distributions, the prediction on the test data with models built on the training data will be inaccurate. One simple solution is to obtain the labeled training data from all data distributions. However, in real-world applications, obtaining labeled training data for all data distribution changes is very difficult if not impossible. In the pedestrian identification example, collecting footstep data under every possible walking speed would not be practical.

Therefore, in this paper, we focus on this common problem: the test data distribution is different from the training set, especially when the difference is significant. However, we have much unlabeled data. In the concept example shown in Figure 1, the unlabeled data marked with darker color have a similar distribution to the labeled data, while the lighter-colored labeled and unlabeled data have different distributions. As the dash-line model shows, if we train a model using either supervised learning or traditional semi-supervised learning, it is not general enough to correctly separate data with a very different distribution from the labeled data. Our ILPC method seeks to attack this problem (see the solid line in Figure 1).

ILPC: Iterative Learning Using Physical Constraints

We propose ILPC to solve inaccuracy caused by distribution changes even for significant changes. ILPC trains a model iteratively and controls the order of unlabeled data used in each iteration according to measurable physical constraints. In this section, we first discuss these two key ideas in details and then present our ILPC algorithm.

Using Multiple Domain Adaptation Models to Cover Gradually Changing Distributions

When data distribution changes significantly, ILPC iteratively constructs multiple domain adaptation models in order to handle a changing distribution. A single domain adaptation model can handle a limited range of distribution changes. In order for the domain adaptation model to predict well with this change, ILPC labels some of the unlabeled data with a prediction confidence score higher than an empirical threshold. In this way, our ILPC method extends the distribution of labeled data. In the next iteration, ILPC constructs another domain adaptation model using also the newly labeled samples, which then predicts a broader data distribution than the previous iteration. Multiple iterations of this extension process will eventually cover all significantly different distributions. The initial domain adaptation model is trained with the initial labeled data distribution and the selected unlabeled data. For example, in pedestrian identification, the initial labeled data are of people's medium walking speeds. When a person walks faster or slower (unlabeled), the training model is extended gradually using both labeled and unlabeled data.

Guiding the Model Distribution Order According to Physical Constraints

To use the iterative approach discussed above, we assume that, within each iteration, the unlabeled and labeled data distributions should not have a significant difference. To meet this assumption, ILPC selects the data for each iteration utilizing the relationship between the data distribution and the physical constraints. First, we define the physical constraints that cause data distribution changes through either heuristic rules or common sense. For example, in the pedestrian identification application, we know from previous research that walking speed and walking pattern (gait) are strongly correlated. Once these constraints are defined, the sensing system extracts the data (feature) along with its corresponding physical constraints. Based on the physical constraint values, the system decides if the distribution of the test data 1) is within the labeled data distribution, 2) has a similar enough distribution to the labeled data to be trained in the next iteration, or 3) has a very different distribution from the labeled data and therefore needs to wait for the model to be extended. We discuss the details of ILPC next.

define the physical constraint x;
uniformly discretize x into $\{x_1, x_2, \cdots, x_k\}$;
for $n \leftarrow 1$ **to** k **do**
  $Data_{selected} = []$;
  for *each sample in* $Data_{unlabeled}$ **do**
    if $x_n < x_{sample} \leq x_{n+1}$ **then**
      $Data_{selected} = Data_{selected} \cup sample$;
    end
  end
  $Prediction, Confidence =$
   $DomainAdaptation(Data_{labeled}, Data_{selected})$;
  $New\_Labeled\_Data = []$;
  for *each* $result$ *in* $Prediction$ **do**
    if $Confidence_{result} > threshold$ **then**
      $New\_Labeled\_Data =$
       $New\_Labeled\_Data \cup$
       $result\ and\ the\ corresponding\ sample$;
    end
  end
  $Data_{labeled} = Data_{labeled} \cup New\_Labeled\_Data$;
end

Algorithm 1: The novel ILPC algorithm.

ILPC Algorithm

Based on the two key ideas introduced above, we present the ILPC algorithm. The pseudo code is shown in Algorithm 1, in which $DomainAdaptation(D_{labeled}, D_{unlabeled})$ is a function that conducts traditional domain adaptation learning (Shi and Sha 2012; Pan et al. 2011). We initialize the system by determining which physical measurements affect the data distribution and then discretize the physical constraints into k levels. We assume that the range of the physical constraint $[x_{min}, x_{max}]$ is uniformly discretized into $\{x_1, x_2, \cdots, x_k\}$, where $x_1 = x_{min}$, $x_k = x_{max}$. That means there are $k-1$ intervals $[x_n, x_{n+1}]$, where $1 \leq n \leq k-1$.

Iterative learning. To address the distribution change in each interval, we construct one domain adaptation model for each interval. Specifically, we construct $k-1$ domain adaptation models. We then assign the n-th model to handle the data distributions subject to the identified physical attribute with values in the range $[x_n, x_{n+1}]$, where $n \in \{1, 2, \cdots, k-1\}$. We determine the value k by observing the empirical histogram of the sample count. We select an appropriate k such that there are sufficiently many data samples in each interval $[x_n, x_{n+1}]$. In the n-th iteration, ILPC constructs the n-th model by domain adaptation methods based on the selected unlabeled data $Data_{selected}$ with attribute val-

ues in $[x_n, x_{n+1}]$ and the labeled data $Data_{labeled}$. Selecting $Data_{selected}$ with attribute values in $[x_n, x_{n+1}]$ guarantees limited distribution change between $Data_{selected}$ and $Data_{labeled}$, which can be handled with high accuracy by existing domain adaptation methods.

Model expansion. Using the n-th model, we label the unlabeled data with high prediction confidence $Data_{selected}$. The final confidence score of a prediction can be calculated from multiple sample points measured for this prediction. In this paper, we introduce two ways to compute a prediction confidence threshold: 1) summing based thresholding and 2) clustering centroid distance based thresholding. The first method is easier to implement but works well when the data of different distributions are not skewed significantly. The second method works on more skewed datasets, though it takes extra calculation.

Experiments

We test ILPC on two real-world sensing datasets to validate that our approach can effectively handle distribution changes between training and test data.

Real-World Sensing Datasets

In this section, we introduce two sensing datasets used in our experiments: the pedestrian identification dataset (Pan et al. 2017) and the mid-earthquake building health monitoring dataset (Xu, Zhang, and Noh 2017).

Pedestrian identification data The floor vibration based gait dataset we use here includes vibration signals from 10 participants who each contribute 10 walking experiments at a number of controlled step frequencies. The controlled step frequencies are μ, $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$, where μ and σ are selected from a prior survey (Öberg, Karsznia, and Öberg 1993). In each walking experiment, the system uses the 7 consecutively detected footstep signals closest to the sensor to classify the subject's identity (Pan et al. 2017). It outputs the majority prediction of these 7 footsteps as the final identity. For the prediction of each footstep, we conduct a one-against-one strategy, i.e., for 10 pedestrians, we conduct a binary classification on $\frac{10\times 9}{2} = 45$ pairs of pedestrians and then average the classification accuracy of the 45 pairs.

Building health monitoring during earthquake data The building vibration based earthquake dataset is collected from 5 buildings with up to 20 floors. Each building contributes 44 samples. For this collected dataset, we predict damage to the structure in terms of Story Drift Ratio (SDR). There are two label classes: no damage ($0\% \leq SDR < 1\%$) and damaged ($SDR \geq 1\%$) (Council 2000). We output the story damage estimation based on the majority vote from multiple data points on the same floor. Each time we make a binary prediction for a story, we have 10 data samples from different sensors. The majority prediction of these 10 data samples gives us our model's final output.

Experimental Setup

This section describes the experimental setup. We first describe the setup of baseline methods. Then we explain our ILPC methods and our methods with random training order. Finally, we detail the parameters of our experiment.

Baseline methods The baseline methods selected are supervised learning SVM (`SVM`) (Chang and Lin 2011), Transductive SVM (`TSVM`) (Joachims 1999b), Graph-based semi-supervised learning (`GSSL`) (Zhu et al. 2003), and Information-Theoretical Learning (`ITL`) (Shi and Sha 2012). We choose them because `TSVM` and `GSSL` are state-of-the-art semi-supervised learning methods and `ITL` outperforms several other domain adaptation methods on our datasets.

Our ILPC methods To demonstrate the benefit of ILPC on datasets with significant distribution change, we integrate the baseline methods into our ILPC algorithm (Algorithm 1), thus creating hybrid methods. We use `TSVM`, `GSSL` and `ITL` as instances of our ILPC method. The ILPC hybrids are therefore `ILPC-TSVM`, `ILPC-GSSL` and `ILPC-ITL`. The detailed parameters of each of these learning methods are the same as the ones used for the baselines.

Random order iterative methods To demonstrate the importance of extending the model in the order guided by measured physical constraints, we further evaluate the iterative learning methods but without the training order. We randomly select the test data from different distributions for each iteration. We refer to these methods as `RandomI-TSVM`, `RandomI-GSSL`, and `RandomI-ITL` corresponding to the selected baseline methods.

Experimental parameter settings The detailed settings of these learning methods are as follows. We use the software SVM^{light} (Joachims 1999a) to run `SVM` and `TSVM` in our experiments with the RBF kernel. Therefore, the parameters that affect these methods include the kernel parameter γ, as well as the weights of training loss introduced by labeled and unlabeled data C_1 and C_2. In our experiments, they are set as $\gamma = 1$, $C_1 = 16$, and $C_2 = 1$ respectively. We select the parameters γ and C_1 through 5-fold cross validation and parameter C_2 as default. In `GSSL`, we use the Euclidean distance as the distance metric in the graph. In `ITL`, the weight of the regularization term is 10, which achieves the highest accuracy according to the accuracy of the validation dataset.

Prediction confidence score We calculate the prediction confidence score based on the output of each method: 1) For `SVM` and `TSVM`, the prediction confidence of a data sample is the distance of this sample to the classification decision boundary. 2) For `GSSL`, the prediction confidence of a data sample is its weight in the constructed graph. 3) For `ITL`, the prediction confidence is the probability output by the sigmoid function in logistic regression.

In the pedestrian identification problem, we calculate the confidence score of one pedestrian as the sum of the footstep model confidence scores within one walking instance. In earthquake building health monitoring, we train a k-means model on the selected unlabeled data. The unlabeled data points closer to the centroids in the k-means model are assigned a higher confidence score. The confidence score of each floor is then calculated as the sum of the confidence scores assigned through clustering analysis.

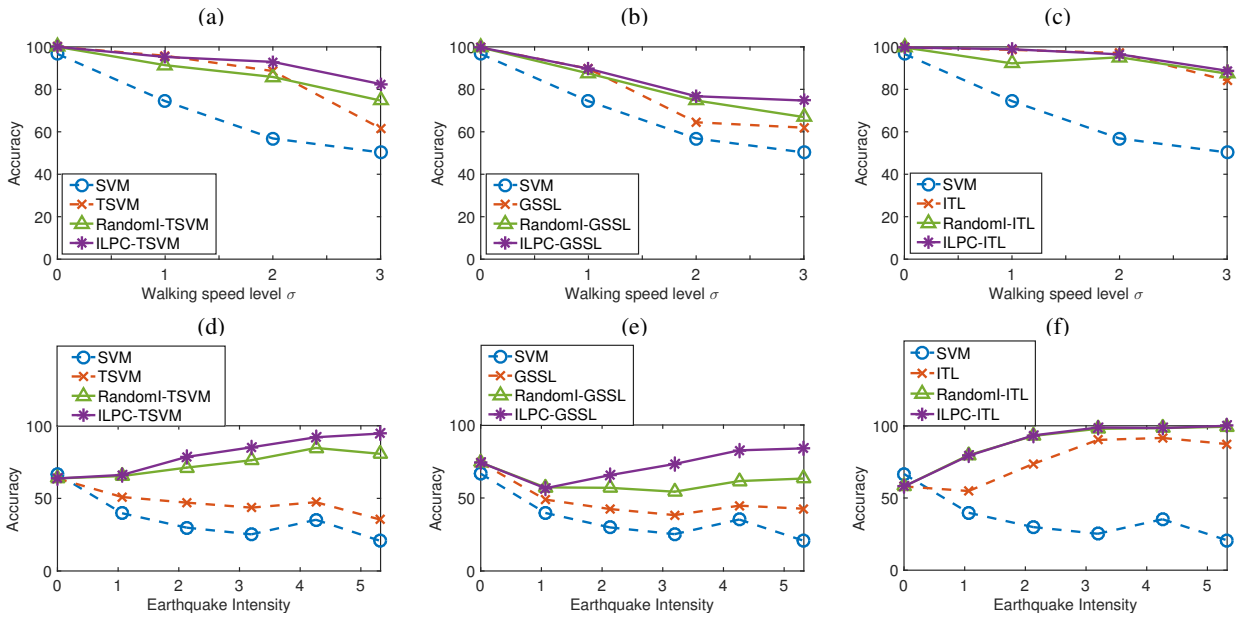

Figure 3: Classification accuracy on changing test data distribution using different learning methods. The x-axis shows the physical attribute value corresponding to the change of data distribution. The y-axis shows the classification accuracy. Figure (a) - (c) show the results of pedestrian identification. Figure (d) - (f) show the results of during-earthquake building health monitoring.

Result Analysis

To evaluate the two key ideas of the ILPC, we first compare traditional domain adaptation methods to our ILPC. Then we compare the results of the algorithm with and without our physical constraint-guided training order.

Evaluation I: Level of Distribution Changes between Training and Test Data.

In Figure 3, we compare the prediction accuracy of our methods (marked as the purple solid line with star markers) and of baseline traditional domain adaptation methods (marked as the red dashed line with cross markers) under different levels of distribution change between training and test data.

Case study 1: pedestrian identification In the case of pedestrian identification, we categorize people's walking speeds into four levels: the average step frequency μ and the gradually increasing step frequencies $\mu + \sigma$, $\mu + 2\sigma$, and $\mu + 3\sigma$, where σ is the standard deviation of a subject's step frequency (Öberg, Karsznia, and Öberg 1993; Pan et al. 2017). Figure 3 (a), (b), and (c) demonstrate identification accuracy for the investigated methods. The x-axis shows 0, σ, 2σ and 3σ, which indicate data samples with the physical constraint value of μ, $\mu + \sigma$, $\mu + 2\sigma$, and $\mu + 3\sigma$ respectively. The y-axis shows the accuracy of predicting the data sample using different methods. As the test data distribution becomes increasingly different from the training data (increasing number of σ), the test accuracy tends to decrease.

Our ILPC methods outperform our baseline domain adaptation methods, especially in cases where test and labeled data have very different walking speeds. `ILPC-TSVM`, `ILPC-GSSL`, and `ILPC-ITL` achieve 92.2%, 85.3%, and 96% average accuracy for all data distribution respectively, while their corresponding traditional domain adaptation methods achieve 86.5%, 79%, and 95%. For all investigated methods, the larger the walking speed difference from the labeled dataset, the lower the identification accuracy. Our ILPC handles training and test data distribution difference better than our baseline methods.

We compared the algorithm runtime of SVM, `TSVM`, and `ILPC-TSVM` in our prior work of pedestrian identification (Pan et al. 2017). Although `TSVM` and `ILPC-TSVM` have significantly higher runtime, compared to collecting the amount of labeled data that covers all data distributions, they still save on the overall time needed.

Case study 2: building health monitoring during earthquakes For this dataset, the labeled data distribution is limited to samples with intensity less than 0.34. First, we test with earthquake data of between 0.34 and 7.64 intensity. We then discretize the earthquake intensity by 0.01, and we present the results in 6 levels to understand the accuracy change trend [1]. Figure 3 (d), (e), and (f) classify the damage estimation results in 6 levels based on earthquake intensity.

Our ILPC methods outperform baseline domain adaptation methods as shown in Figure 3 (d), (e), and (f). At the maximum data distribution difference, the `TSVM`, `GSSL`, and `ITL` achieve an accuracy of 35.3%, 42.7%, and 87.3%, respectively, while our ILPC methods `ILPC-TSVM`, `ILPC-GSSL`, and `ILPC-ITL` achieve 94.7%, 84%, and 99.3%, resulting in up to $2.7\times$ improvement compared to traditional domain adaptation methods. The average accuracy over all data distributions is 48%, 48.6%, and 75.9% respectively for the baseline meth-

[1] In Figure 3, we only show up to 5.2 due to this discretization

ods and 80.1%, 72.8%, and 88% for ILPC methods, which shows up to a $1.7\times$ improvement.

The accuracy improvement by ILPC is caused by labeling test data in each iteration. As the iteration and labeling continue, more and more samples with features that appear in high-intensity earthquakes are labeled. Our ILPC methods show an increasing trend in prediction accuracy when the earthquake intensity increases. This is because degree of building damage changes at different intensity levels. The data distribution of high-intensity earthquakes are more different from those of low-intensity earthquakes, resulting in a higher classification accuracy.

Evaluation II: The Impact of Iteration Order

To understand the importance of physical constraints in guiding the model's distribution order, we focus on the comparison between the random learning order (marked as the green solid line with triangle markers) and the physical constraints guided learning order (our ILPC methods, marked as the purple solid line with star markers) in Figure 3. In both pedestrian identification and building health monitoring applications, the accuracy of our ILPC methods is higher than that of the random-order methods. For `ILPC-TSVM` and `ILPC-GSSL`, the larger the distribution difference is, the better our ILPC methods perform.

For the pedestrian identification application, we show prediction results in Figure 3 (a), (b) and (c). When the distribution difference is 3σ, the `ILPC-TSVM` and `ILPC-GSSL` achieves 82.5% and 74.7% accuracy, while the random order version achieves only 74.7% and 66.9% accuracy. This happens because using physical constraints to control the order limits the data distribution change in each iteration. Therefore, more test data can be labeled with high confidence and contribute to the later iterations as labeled data. `RandomI-ITL` performs worse than `ILPC-ITL` at the beginning of the model extension (1σ). It achieves 92.2% while `ILPC-ITL` achieve 98.9%. When more data is labeled through more iterations, the accuracy increases.

For the building damage estimation application, we show prediction results in Figure 3 (d), (e) and (f). `ILPC-TSVM` and `ILPC-GSSL` achieve 94.7% and 84% accuracy when predicting building damage during the highest intensity earthquakes, higher than that of `RandomI-TSVM` and `RandomI-GSSL` (80.7% and 63.3%). `RandomI-ITL` and `ILPC-ITL` show up to 99.3% and 100% accuracy in this dataset, and therefore `ILPC-ITL` does not show significant improvement on `RandomI-ITL`. `RandomI-ITL` outperforms `ITL` due to the iteration and labeling. `ILPC-ITL` achieves a similar level of accuracy to `RandomI-ITL`, because `ITL` is robust to the data distribution change in this dataset. Using iteration (`RandomI-ITL` and `ILPC-ITL`) amplifies its adaptation ability. Using physical constraints to control distribution order consistently improves the accuracy of our methods.

Related Works

ILPC is closely related to and built upon two types of learning methods: transductive learning and domain adaptation.

Transductive learning is a technique that addresses the distribution change between training and test data, including self-training, low-density separation, and graph-based methods (Chapelle, Schölkopf, and Zien 2010). Low-density separation methods, such as Transductive SVM (`TSVM`) (Joachims 1999b; Chapelle, Schölkopf, and Zien 2010), tend to place decision boundaries where the unlabeled data has low density. The graph-based methods (`GSSL`) (Zhu et al. 2003) construct a graph where the nodes are the labeled and unlabeled data points. In addition to having good accuracy in the labeled training data, the graph-based methods learn a model where the prediction of unlabeled data is smooth in the constructed graph. Previous works (Belkin and Niyogi 2004; Li and Zhou 2011; Li et al. 2013; Wang, Wang, and Li 2016; Li, Wang, and Zhou 2016) show that traditional transductive learning methods such as `TSVM` and `GSSL`, may have decreased accuracy with unlabeled data in the model training. To avoid accuracy decrease, unlabeled instances are selectively used in the transductive learning (Li and Zhou 2011; Wang, Wang, and Li 2016; Li, Wang, and Zhou 2016).

Domain adaptation adapts a model learned from data in a source domain to data with a different distribution in a target domain. There are many domain adaptation algorithms developed, such as Transfer Component Analysis (TCA) (Pan et al. 2011), Maximum Independence Domain Adaptation (MIDA) (Yan, Kou, and Zhang 2016), Subspace Alignment (SA) (Fernando et al. 2013), Information-Theoretical Learning (ITL) (Shi and Sha 2012), Geodesic flow kernel (GFK) (Gong et al. 2012) and Stationary Subspace Analysis (SSA) (Von Bünau et al. 2009). In general, both transductive learning and domain adaptation methods can only handle distribution change in a limited range. Our ILPC methods can be used with both of them to handle a larger range of distribution changes.

Conclusion

In physical systems, the training and test data distributions are often significantly different, resulting in potentially inaccurate learning models. Previous domain adaptation methods can improve accuracy but are limited by the distribution change range. We develop ILPC, an iterative learning method guided by sensible physical constraints that indicate the data distribution similarity. First, we train multiple domain adaptation models iteratively to cover different parts of the gradually changing distribution. Second, in each iteration, we use the trained domain adaptation model to label the test data and extend the model. The test data trained in each iteration is guided by the assigned physical constraints to ensure the similarity of the data distribution handled in each iteration. We evaluate our ILPC on two real-world sensing datasets and it shows up to $2.7\times$ prediction accuracy improvement compared to the corresponding traditional domain adaptation.

Acknowledgements

This work is partially supported by NSF (CNS-1149611, CMMI-1653550 and 1344768), Intel, and Google.

References

[Belkin and Niyogi 2004] Belkin, M., and Niyogi, P. 2004. Semi-supervised learning on Riemannian manifolds. *Machine learning* 56(1-3):209–239.

[Chang and Lin 2011] Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

[Chapelle, Schölkopf, and Zien 2010] Chapelle, O.; Schölkopf, B.; and Zien, A. 2010. Semi-supervised learning. *IEEE Transactions on Neural Networks* 20(3):542–542.

[Council 2000] Council, B. S. S. 2000. Prestandard and commentary for the seismic rehabilitation of buildings. *Report FEMA-356, Washington, DC*.

[Fernando et al. 2013] Fernando, B.; Habrard, A.; Sebban, M.; and Tuytelaars, T. 2013. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, 2960–2967.

[Gong et al. 2012] Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, 2066–2073. IEEE.

[Gubbi et al. 2013] Gubbi, J.; Buyya, R.; Marusic, S.; and Palaniswami, M. 2013. Internet of things (iot): A vision, architectural elements, and future directions. *Future generation computer systems* 29(7):1645–1660.

[Joachims 1999a] Joachims, T. 1999a. Making large-scale SVM learning practical. In Schölkopf, B.; Burges, C.; and Smola, A., eds., *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press. chapter 11, 169–184.

[Joachims 1999b] Joachims, T. 1999b. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 200–209. Morgan Kaufmann Publishers Inc.

[Li and Zhou 2011] Li, Y.-F., and Zhou, Z.-H. 2011. Improving semi-supervised support vector machines through unlabeled instances selection. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 386–391. AAAI Press.

[Li et al. 2013] Li, Y.-F.; Tsang, I. W.; Kwok, J. T.; and Zhou, Z.-H. 2013. Convex and scalable weakly labeled svms. *Journal of Machine Learning Research* 14(1):2151–2188.

[Li, Wang, and Zhou 2016] Li, Y.; Wang, S.; and Zhou, Z. 2016. Graph quality judgement: a large margin expedition. *Proceedings of the 25th International Joint Confernece on Artificial Intelligence, New York, NY*.

[Notion Inc. 2017] Notion Inc. 2017. *Home awareness, simplified. Monitor your home with a single sensor, wherever you are.* `http://getnotion.com/`.

[Öberg, Karsznia, and Öberg 1993] Öberg, T.; Karsznia, A.; and Öberg, K. 1993. Basic gait parameters: reference data for normal subjects, 10-79 years of age. *Journal of rehabilitation research and development* 30:210–210.

[Pan et al. 2011] Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

[Pan et al. 2015] Pan, S.; Wang, N.; Qian, Y.; Velibeyoglu, I.; Noh, H.; and Zhang, P. 2015. Indoor person identification through footstep induced structural vibration. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*.

[Pan et al. 2017] Pan, S.; Yu, T.; Mirshekari, M.; Fagert, J.; Bonde, A.; Mengshoel, O. J.; Noh, H. Y.; and Zhang, P. 2017. Footprintid: Indoor pedestrian identification through ambient structural vibration sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3(89):31.

[Samsung Inc. 2017] Samsung Inc. 2017. *The easiest way to turn your home into a smart home.* `https://www.samsung.com/us/smart-home/smartthings/`.

[Shi and Sha 2012] Shi, Y., and Sha, F. 2012. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 1079–1086.

[Von Bünau et al. 2009] Von Bünau, P.; Meinecke, F. C.; Király, F. C.; and Müller, K.-R. 2009. Finding stationary subspaces in multivariate time series. *Physical review letters* 103(21):214101.

[Wang, Wang, and Li 2016] Wang, H.; Wang, S.-B.; and Li, Y.-F. 2016. Instance selection method for improving graph-based semi-supervised learning. In *Pacific Rim International Conference on Artificial Intelligence*, 565–573. Springer.

[Xu, Zhang, and Noh 2017] Xu, S.; Zhang, P.; and Noh, H. Y. 2017. Information theoretic approach for seismic damage localization. In *Proceedings of the 16th World Conference on Earthquake (16WCEE)*. WCEE.

[Yan, Kou, and Zhang 2016] Yan, K.; Kou, L.; and Zhang, D. 2016. Domain adaptation via maximum independence of domain features. *arXiv preprint arXiv:1603.04535*.

[Zhu et al. 2003] Zhu, X.; Ghahramani, Z.; Lafferty, J.; et al. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, 912–919.