
Coping with New User Problems: Transfer Learning in Accelerometer-Based Human Activity Recognition

Hiroataka Hachiya Masashi Sugiyama
Tokyo Institute of Technology
{hachiya@sg. sugi@}cs.titech.ac.jp

Naonori Ueda
NTT Communication Science Laboratories
ueda@cslab.kecl.ntt.co.jp

1 Introduction

Human activity recognition from accelerometric data (e.g., obtained by an iPhone) is gathering a great deal of attention recently since it can be used for various purposes such as diagnostic and therapeutic decision making. Standard classification algorithms such as the support vector machines and logistic regression are useful for building an accurate recognition system, but gathering labeled training data is often bothersome particularly for new users. Thus, transfer learning from a large number of existing users is necessary to enhance the usability of activity recognition systems. In this work, we demonstrate that importance-sampling-based transfer learning highly improves the accuracy through real-world human activity recognition experiments. Key ingredients of our method are covariate shift adaptation for consistent parameter estimation and computationally efficient probabilistic classification for large-scale training and prediction confidence acquisition.

2 Transfer Learning via Covariate Shift Adaptation

Supervised Learning under Covariate Shift: First, let us formulate the supervised learning problem¹. Let $\{(\mathbf{x}_n^{\text{tr}}, y_n^{\text{tr}})\}_{n=1}^{N_{\text{tr}}}$ be the training samples where \mathbf{x}_n^{tr} is a training input point drawn from a probability density $p_{\text{tr}}(\mathbf{x})$ and y_n^{tr} is a training output value following a conditional probability density $p^*(y|\mathbf{x} = \mathbf{x}_n^{\text{tr}})$. Let $(\mathbf{x}^{\text{te}}, y^{\text{te}})$ be a test sample where \mathbf{x}^{te} is a test input point following a probability density $p_{\text{te}}(\mathbf{x})$ and y^{te} is a test output value following $p^*(y|\mathbf{x} = \mathbf{x}^{\text{te}})$. Note that the test sample is not available in the training phase, but will be provided in the test phase.

The goal of supervised learning is to obtain an approximation $\hat{f}(\mathbf{x})$ that minimizes the generalization error G (or the expected test error):

$$G \equiv \iint \text{loss}(\hat{f}(\mathbf{x}), y) p^*(y|\mathbf{x}) p_{\text{te}}(\mathbf{x}) d\mathbf{x},$$

where $\text{loss}(\hat{y}, y)$ is the *loss* function which measures the discrepancy between the true output value y and its estimate \hat{y} . A standard method to learn the parameter θ in the model $f(\mathbf{x}; \theta)$ would be *empirical risk minimization* (ERM):

$$\hat{\theta}_{\text{ERM}} \equiv \arg \min_{\theta} \left[\frac{1}{N_{\text{tr}}} \sum_{n=1}^{N_{\text{tr}}} \text{loss}(f(\mathbf{x}_n^{\text{tr}}; \theta), y_n^{\text{tr}}) \right].$$

If training and test input points follow the same probability densities, i.e., $p_{\text{te}}(\mathbf{x}) = p_{\text{tr}}(\mathbf{x})$, $\hat{\theta}_{\text{ERM}}$ converges to the optimal parameter θ^* , where $\theta^* \equiv \arg \min_{\theta} [G]$. However, in many real-world problems, the assumption $p_{\text{tr}}(\mathbf{x}) = p_{\text{te}}(\mathbf{x})$ is often violated. For example, in human activity recognition, distributions of accelerometric data tend to be different depending on users since measurement conditions such as the position and orientation of the sensor are diverse.

Situations where training and test input points follow different probability densities, e.g., $p_{\text{tr}}(\mathbf{x}) \neq p_{\text{te}}(\mathbf{x})$, but the conditional density $p^*(y|\mathbf{x})$ of output values given input points is unchanged is

¹Effectively, an input \mathbf{x} and an output y represent an accelerometer sample and its corresponding activity, respectively. $p_{\text{te}}(\mathbf{x})$ and $p_{\text{tr}}(\mathbf{x})$ represent the probability densities of a new user (for whom we want to design a activity recognize) and other exiting users, respectively. See Section 3 for details.

called *covariate shift* (Shimodaira, 2000). Under covariate shift, most of the standard learning techniques do not work well due to differing distributions. More specifically, $\hat{\boldsymbol{\theta}}_{\text{ERM}}$ does not converge to $\boldsymbol{\theta}^*$ if the model $f(\mathbf{x}; \boldsymbol{\theta})$ is not correctly specified (i.e., the ‘true’ function is not included in the model). Since the correctness of the model is not usually guaranteed in real-world problems, the inconsistency of ERM is critical in practice.

Importance sampling is a standard technique to compensate for the difference of densities using importance weight $w^*(\mathbf{x}) = p_{\text{te}}(\mathbf{x})/p_{\text{tr}}(\mathbf{x})$ as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{IWERM}} &\equiv \arg \min_{\boldsymbol{\theta}} \left[\frac{1}{N_{\text{tr}}} \sum_{n=1}^{N_{\text{tr}}} w^*(\mathbf{x}_n^{\text{tr}}) \text{loss}(f(\mathbf{x}_n^{\text{tr}}; \boldsymbol{\theta}), y_n^{\text{tr}}) \right] \\ &\xrightarrow{N_{\text{tr}} \rightarrow \infty} \arg \min_{\boldsymbol{\theta}} \left[\iint w^*(\mathbf{x}) \text{loss}(f(\mathbf{x}; \boldsymbol{\theta}), y) p^*(y|\mathbf{x}) p_{\text{tr}}(\mathbf{x}) d\mathbf{x} dy \right] = \arg \min_{\boldsymbol{\theta}} [G] = \boldsymbol{\theta}^*. \end{aligned}$$

This shows that the solution of *importance-weighted ERM* (IWERM) $\hat{\boldsymbol{\theta}}_{\text{IWERM}}$ converges to the optimal solution $\boldsymbol{\theta}^*$ even under covariate shift with misspecified models.

However, the importance weight $w^*(\mathbf{x}_n^{\text{tr}})$ is unknown in practice, and thus we cannot use the importance weighting technique directly.

Importance Weight Estimation: For importance estimation, we use *unconstrained least-squares importance fitting* (uLSIF) (Kanamori et al., 2009). The main idea of uLSIF is to directly estimate the ratio of two densities $w^*(\mathbf{x}) \equiv p_{\text{te}}(\mathbf{x})/p_{\text{tr}}(\mathbf{x})$ using least-squares fitting without estimating $p_{\text{tr}}(\mathbf{x})$ and $p_{\text{te}}(\mathbf{x})$. Here we assume that unlabeled test input samples $\{\mathbf{x}_n^{\text{te}}\}_{n=1}^{N_{\text{te}}}$ independently drawn from $p_{\text{te}}(\mathbf{x})$ are available (i.e., *semi-supervised learning*).

Let us model the importance $w^*(\mathbf{x})$ by $w(\mathbf{x}; \boldsymbol{\alpha}) = \phi(\mathbf{x})^\top \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a parameter vector and $\phi(\mathbf{x})$ is a basis function vector. We determine the parameters $\boldsymbol{\alpha}$ so that the following squared error J is minimized:

$$\begin{aligned} J(\boldsymbol{\alpha}) &\equiv \frac{1}{2} \int \left(w(\mathbf{x}; \boldsymbol{\alpha}) - w^*(\mathbf{x}) \right)^2 p_{\text{tr}}(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{2} \int w(\mathbf{x}; \boldsymbol{\alpha})^2 p_{\text{tr}}(\mathbf{x}) d\mathbf{x} - \int w(\mathbf{x}; \boldsymbol{\alpha}) p_{\text{te}}(\mathbf{x}) d\mathbf{x} + \text{Const}. \end{aligned}$$

By empirical approximation and regularization, we obtain the following optimization problem:

$$\hat{\boldsymbol{\alpha}} \equiv \arg \min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{H} \boldsymbol{\alpha} - \mathbf{h}^\top \boldsymbol{\alpha} + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right],$$

where $\frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha}$ ($\lambda \geq 0$) is a regularization term, $\mathbf{H} \equiv \frac{1}{N_{\text{tr}}} \sum_{n=1}^{N_{\text{tr}}} \phi(\mathbf{x}_n^{\text{tr}}) \phi(\mathbf{x}_n^{\text{tr}})^\top$, and $\mathbf{h} \equiv \frac{1}{N_{\text{te}}} \sum_{n=1}^{N_{\text{te}}} \phi(\mathbf{x}_n^{\text{te}})$. Then the uLSIF solution can be analytically computed as $\hat{\boldsymbol{\alpha}} = (\mathbf{H} + \lambda \mathbf{I})^{-1} \mathbf{h}$, where \mathbf{I} denotes the identity matrix. Since the importance weight is non-negative by definition, we modify the solution as $\hat{w}(\mathbf{x}) \equiv \max(0, \phi(\mathbf{x})^\top \hat{\boldsymbol{\alpha}})$.

Least-Squares Probabilistic Classifier: In practical human activity recognition, obtaining a *confidence* of prediction is useful because prediction with low confidence can be rejected. To this end, kernel logistic regression for estimating the class-posterior probability $p^*(y|\mathbf{x})$ is useful. However, although sophisticated optimization toolboxes of, e.g., (quasi-)Newton methods are readily available, training a large-scale kernel logistic model is still challenging. Here we propose to use a probabilistic classifier called the *least-squares probabilistic classifier* (LSPC) (Sugiyama, 2010), which is a computationally efficient alternative to kernel logistic regression.

The goal of LSPC is to estimate the class-posterior $p^*(y|\mathbf{x})$ from training samples $\{(\mathbf{x}_n^{\text{tr}}, y_n^{\text{tr}})\}_{n=1}^{N_{\text{tr}}}$. Let us model the class posterior probability $p^*(y|\mathbf{x})$ for class y by $p(y|\mathbf{x}; \boldsymbol{\theta}^{(y)}) \equiv \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\theta}^{(y)}$, where $\boldsymbol{\theta}^{(y)}$ is a parameter vector and $\boldsymbol{\psi}(\mathbf{x})$ is a basis function vector. We determine the parameter

$\theta^{(y)}$ so that the following squared error J_y is minimized:

$$\begin{aligned} J_y(\theta^{(y)}) &\equiv \frac{1}{2} \int \left(p(y|\mathbf{x}; \theta^{(y)}) - p^*(y|\mathbf{x}) \right)^2 p_{te}(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{2} \int p(y|\mathbf{x}; \theta^{(y)})^2 p_{te}(\mathbf{x}) d\mathbf{x} - \int p(y|\mathbf{x}; \theta^{(y)}) p^*(y|\mathbf{x}) p_{te}(\mathbf{x}) d\mathbf{x} + \text{Const.} \end{aligned}$$

By importance-weighted empirical approximation and regularization, we obtain the following optimization problem:

$$\hat{\theta}^{(y)} \equiv \arg \min_{\theta} \left[\frac{1}{2} \theta^\top \mathbf{Q} \theta - \mathbf{q}_y^\top \theta + \frac{\gamma}{2} \theta^\top \theta \right],$$

where $\frac{\gamma}{2} \theta^\top \theta$ is a regularization term, $\mathbf{Q} \equiv \frac{1}{N_{tr}} \sum_{n=1}^{N_{tr}} w^*(\mathbf{x}_n^{tr}) \psi(\mathbf{x}_n^{tr}) \psi(\mathbf{x}_n^{tr})^\top$, and $\mathbf{q}_y \equiv \frac{1}{N_{tr}} \sum_{n: y_n^{tr}=y} w^*(\mathbf{x}_n^{tr}) \psi(\mathbf{x}_n^{tr})$. The LSPC solution is given analytically as $\hat{\theta}^{(y)} = (\mathbf{Q} + \gamma \mathbf{I})^{-1} \mathbf{q}_y$. Since the class-posterior probability is non-negative by definition, we modify the solution as follows (Yamada et al., 2010): $\hat{p}(y|\mathbf{x}) \equiv \max(0, \psi(\mathbf{x})^\top \hat{\theta}^{(y)}) / Z$ if $Z = \sum_{y=1}^c \max(0, \psi(\mathbf{x})^\top \hat{\theta}^{(y)}) > 0$; otherwise $\hat{p}(y|\mathbf{x}) \equiv 1/c$ where c denotes the number of classes.

Thanks to the analytic-form solution, LSPC is computationally much more efficient than kernel logistic regression². Nevertheless, LSPC was demonstrated to have comparable classification accuracy to kernel logistic regression (Sugiyama, 2010; Yamada et al., 2010).

Importance-Weighted Cross-Validation: *Cross-validation* (CV) is a standard model selection method for choosing the value of tuning parameters such as basis parameters (e.g., the Gaussian kernel width) and regularization parameter γ . However, under covariate shift, ordinary CV is highly biased due to differing distributions. To cope with this problem, a variant of CV called *importance-weighted CV* (IWCV) has been proposed in Sugiyama et al. (2007).

Let us randomly divide the training set into $\mathcal{D} = \{(\mathbf{x}_n^{tr}, y_n^{tr})\}_{n=1}^{N_{tr}}$ into K disjoint non-empty subsets $\{\mathcal{D}_k\}_{k=1}^K$ of (approximately) the same size. Let $f_k(\mathbf{x})$ be a function learned from $\mathcal{D} \setminus \mathcal{D}_k$ (i.e., without \mathcal{D}_k). Then the k -fold IWCV estimate of the generalization error G is given by

$$\hat{G}_{\text{IWCV}} = \frac{1}{K} \sum_{k=1}^K \frac{1}{|\mathcal{D}_k|} \sum_{(\mathbf{x}^{tr}, y^{tr}) \in \mathcal{D}_k} w^*(\mathbf{x}) \text{loss}(f_k(\mathbf{x}^{tr}), y^{tr}),$$

where $|\mathcal{D}_k|$ is the number of samples in the subset \mathcal{D}_k . It was proved that IWCV gives an almost unbiased estimate of the generalization error even under covariate shift (Sugiyama et al., 2007).

3 Experiments

In this section, we investigate the performance of the proposed method in real-world human activity recognition. We use three-axis accelerometric data collected by iPodTouch available at <http://alkan.mns.kyutech.ac.jp/web/data.html>. In the data collection procedure, subjects were asked to perform a specific task such as walking, running, and going up the stairs, and its accelerometric data were recorded by iPodTouch. The duration of each task was arbitrary and the sampling rate was 20 Hz with small variations.

To extract features from the accelerometric data, each data-stream was segmented in a sliding window manner with window width 5 seconds and sliding step 1 second. Depending on subjects, the position and orientation of iPodTouch was arbitrary—held by hand or kept in a pocket or a bag. For this reason, we decided to take the ℓ_2 -norm of the 3-dimensional acceleration vector at each time step, and computed the following 5 features from each window: *mean*, *standard deviation*, *fluctuation of amplitude*, *average energy* and *frequency-domain entropy* (Bao & Intille, 2004; Bharatula et al., 2005). Note that these features are orientation invariant.

Let us consider a situation where a new user wants to use the activity recognition system. However, he/she does not want to label their accelerometric data. Thus, there is no labeled sample for the new

²Computing the LSPC solution corresponding to a single Newton step (i.e., iteratively-weighted least-squares) of kernel logistic regression.

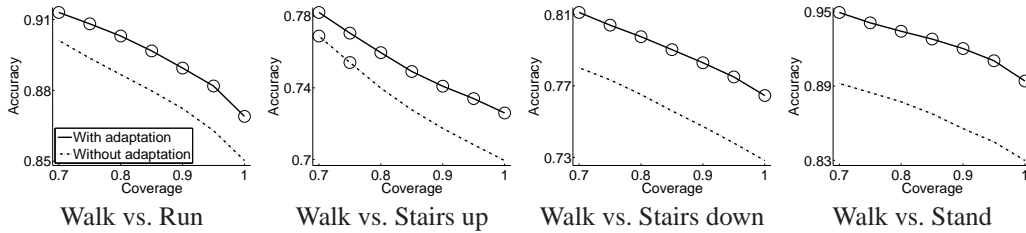


Figure 1: Accuracy versus coverage averaged over 10 new users and 10 trials. There are four two-class classification problems to recognize walking or running, walking or going stairs up, walking or going stairs down and walking or standing. The symbol ‘o’ indicates the fact that the method is the best or comparable to the best one, determined by the t -test at the significance level 5%.

user. On the other hand, a large number of unlabeled samples for the new user as well as a large number of labeled data for existing users are available.

Let $\mathcal{D}_{\text{own}}^{\text{unlab}}$ be sets of 1000 unlabeled accelerometric data of a new user and $\mathcal{D}_{\text{others}}$ be the set labeled accelerometric data for 9 existing users. Each existing user has 100 labeled samples for each action. We use 400 Gaussian kernels as basis functions $\psi(x)$ for modeling the class-posterior model $p(y|x; \theta^{(y)})$, where Gaussian centers were chosen from the set of unlabeled samples $\mathcal{D}_{\text{own}}^{\text{unlab}}$. The proposed method (IWLSPC+IWCV+uLSIF, where estimated importance weights were averaged for each user) is compared with a no-adaptation baseline (LSPC+CV, where data from the existing users are directly used for training the new user’s classifier).

Figure 1 depicts the prediction accuracies for 1000 test samples averaged over 100 cases (10 new users and 10 trials per user) with varying coverage values. The coverage is the ratio of test sample size used for the evaluation. For example, the coverage 0.8 indicates that 20% of test samples with lower prediction confidence based on the predicted class-posterior probability are not used when evaluating the accuracy. The graphs show that the proposed method significantly outperforms the baseline method over the entire coverage. In addition, the accuracy monotonically increases as the coverage decreases. This implies that the prediction confidence helps us to further improve the performance of our proposed method.

4 Conclusions

We proposed to use an importance-sampling-based transfer learning technique for coping with the new user problem in human activity recognition. Experiments with real-world data illustrated the usefulness of the proposed method. The proposed method is computationally highly efficient and scalable to massive datasets. We are currently preparing data collected from hundreds of users for a year, and our future work will apply the proposed method to large-scale knowledge transfer.

This work was supported by the FIRST program.

References

- Bao, L., & Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. *Proc. 2nd IEEE International Conference on Pervasive Computing* (pp. 1–17).
- Bharatula, N. B., Stager, M., Lukowicz, P., & Troster, G. (2005). Empirical study of design choices in multi-sensor context recognition systems. *Proc. 2nd Intl. Forum on Applied Wearable Computing* (pp. 79–93).
- Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10, 1391–1445.
- Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 227–244.
- Sugiyama, M. (2010). Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, E93-D, 2690–2701.
- Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8, 985–1005.
- Yamada, M., Sugiyama, M., Wichern, G., & Simm, J. (2010). *Improving the accuracy of least-squares probabilistic classifiers* (Technical Report IBISML2010-32). IEICE Technical Report.