Machine Learning Project Report

Transfer Learning based Activity Recognition via Domain Adaptation

Jindong Wang, 201418013229092
Liping Jia, 201518013229093
{wangjindong, jialiping}@ict.ac.cn

July 3, 2016

1 Introduction

Research on transfer learning based activity recognition is on the go. Prevalent works seek the potential of transferring existing knowledge to the target domain through instance based transfer, feature based transfer, as well as parameter based transfer methods respectively. For more information, please refer to D. Cook's survey [1]. And for background of transfer learning, please see Pan and Yang's survey [3].

In our work, we propose to use transfer learning to implement our activity recognition task based on transfer component analysis (TCA) [2]. TCA tries to learn some transfer components across domains in a reproducing kernel Hilbert space using maximum mean discrepancy. After that we can utilize some machine learning methods to perform classification. Detailed information on problem formulation, method, and experiment is coming as follows.

2 Problem Formulation

In the context of activity recognition, there are several dimensional sensor data that can be used to perform recognition. We take notice of most distinguishable ones: accelerometer and gyroscope.

Let's say we have fully labeled gyroscope readings available, while we want to annotate the unknown accelerometer from the same person while performing the same activity at the same time.

Input:labeled $\mathbf{D}_S = \{\mathbf{X}_i, y_i\}_{i=1}^m$, and unlabeled $\mathbf{D}_T = \{\mathbf{X}_i\}_{i=1}^t$, while $y_i \in \{1, 2, \cdots, C\}$ is the label.

Output: the predicted value for $\mathbf{X}_i$ in $\mathbf{D}_T$: y_is.

Method:Transfer learning using TCA. After that, we train a model using labeled new $\mathbf{X}_{src}$ and y_{src} and use that to validate the new $\mathbf{X}_{tar}$.

3 Method

The distance between a source and target domain can be calculated using MMD: $\text{Dist}(X_S, X_T) = \|\frac{1}{n_1}\sum_{i=1}^{n_1}\phi(X_{S_i}) - \frac{1}{n_2}\sum_{i=1}^{n_2}\phi(X_{Tc_i})\|_{\mathcal{H}}^2$, where ϕ is the feature map induced by a universal kernel.

After applying MMD [2], let K be the kernel matrix, $K_{S,S}, K_{S,T}, K_{T,T}$ denote the source domain, cross-domain and target domain data respectively, then

$$K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$$

Therefore, the distance between the transformed source and target domain becomes

$$\text{dist}(\mathbf{X}'_{src}, \mathbf{X}'_{tar}) = \text{trace}(KL)$$

where

$$L = \begin{cases} \frac{1}{n_1^2} & x_i, x_j \in \mathbf{X}_{src} \\ \frac{1}{n_2^2} & x_i, x_j \in \mathbf{X}_{tar} \\ -\frac{1}{n_1 n_2} & \text{otherwise} \end{cases}$$

Furthermore, the problem can be formulated as

$$\begin{aligned} \min \quad & \text{trace}(KL) - \lambda \text{trace}(K) \\ \text{s.t.} \quad & K_{ii} + K_{jj} - 2K_{ij} + 2\epsilon = d_{ij}^2 \\ & K\mathbf{1} = -\epsilon\mathbf{1} \end{aligned}$$

Which is an semidefinite programming problem.

In order to fit this problem to the standard SDP form, we did the following work:

$$\begin{aligned} \min \quad & \text{trace}((L - \lambda I)K) \\ \text{s.t.} \quad & A^{(m)} \bullet K = D_{ij} \end{aligned}$$

where

$$A^{(m)} = \begin{cases} A_{ii}^{(m)} = A_{jj}^{(m)} = 1 \\ A_{ij}^{(m)} = A_{ji}^{(m)} = -1 \end{cases}$$

and

$$C \bullet X := \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} X_{ij} = \text{trace}(CX)$$

However, it it extremely burdensome to solve an SDP. Actually the time complexity is $O(n_1 + n_2)^{6.5}$! So we continue to shift our problem to TCA, which is transfer component analysis.

Our problem fits the unsupervised version of TCA. Let $H = I_{n_1+n_2} - (\frac{1}{n_1+n_2})\mathbf{1}\mathbf{1}^T$, and $W = K^{-1/2}\widetilde{W}$ where $\widetilde{W} \in \mathbb{R}^{(n_1+n_2)\times m}$ transforms the empirical kernel map features to an m-dimensional space.

Our learning problem becomes:

$$\begin{aligned} \min_{W} \quad & \text{tr}(W^T KLKW) + \mu \text{tr}(W^T W) \\ \text{s.t.} \quad & W^T KHKW = I_m \end{aligned}$$

Where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix and $\mu > 0$.

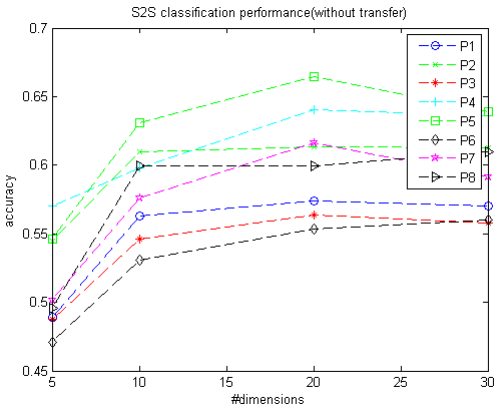
After TCA, there will be new representations for the source and target data, where their distance is minimized while preserving their inner structural information. Since the new source and target data are in the same feature space, we can directly apply traditional machine learning techniques. Then we can train a machine learning model on the new source data, and test it on the new target data.

4 Experiment

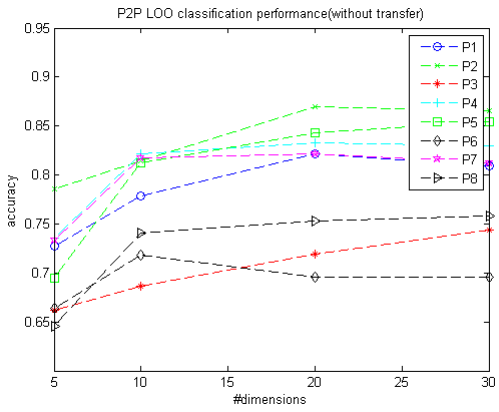
In this section, we perform 3 kinds of experiment to evaluate the feasibility of TCA. The 3 kinds of experiment include: 1) basic classification without transfer; 2) P2P transfer: transfer from person to person for the same feature spaces and 3) S2S transfer: transfer from sensor to sensor for the same person.

We exploit an activity recognition dataset from UCI [1]. The dataset contains activities of 8 people performing 19 activities. Contributors explored the sensors on 5 different body parts with 3-axial accelerometer, gyroscope and magnetometer on each part respectively. That led to 45 sensor readings per frame (3 * 5 * 3).

[1] `http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities`



(a) No transfer-1v1



(b) No transfer-LOO

Figure 1: Experimental result of person to person classification without transfer. (a) is 1v1 no transfer results, using 1 person to train the model and the other to test and (b) is LOO test, using 7 persons to train the model and the other one to test. Y axis represents the average accuracy for every person, and X axis represents the dimensions after reduction. The next figures use the same axis.

Features are extracted for every axis. For every sensor, fist we combine the 3 axis together using $a = \sqrt{x^2 + y^2 + z^2}$. We exploit sliding window technique to extract features (window size = 5s). Then we extract 27 features from both time and frequency domain, leading to 405 dimensions in total (27 * 3 * 5 = 405). There are 142500 rows of data in the original dataset, while after feature extraction, there are 9120 rows. After feature extraction, we normalize the data of every column into $[0, 1]$.

The most important part in our experiment is TCA, which we will introduce in the sequel. After TCA, traditional machine learning task has to be done. We simply choose random forest as the classifier to train a model, and then use this model to test the unlabeled target data. Accuracy metric is used as the evaluation metric.

The 3 kinds of experiments are as follows.

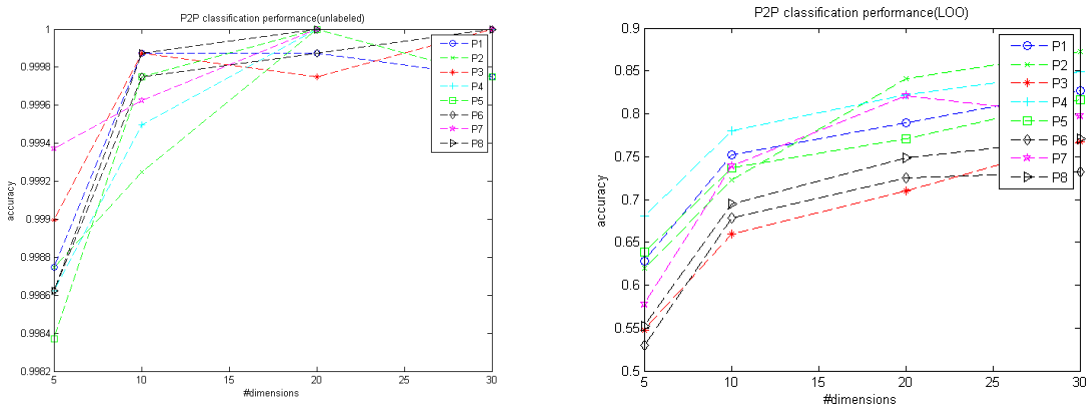
4.1 Basic classification without transfer

This experiment aims to investigate the possibility of learning a classification model without transfer. For every person, we train a model on his solo data using random forest, and then we test our model using other 7 persons' data. In order to act as comparison, we perform PCA to the data to reduce the dimensionality to 5, 10, 20, 30 respectively. The results are as Figure 1(a). On the other hand, in order to evaluate the full potential of non-transfer, we also adopt an LOO (leave one out) experiment, i.e. we train a model using 7 persons' data and test the model on the other one. The experimental result is as Figure 1(b)

Here the accuracy of P1 means the test result of P1, either using other 7 person's data as training set (LOO), or using other 7 persons one by one as the training set and test on P1 respectively (1v1, here accuracy is the mean of 7). From the results, we can see that traditional machine learning method does not generate satisfying results when faced with different users. As the dimensionality after reduction decreases from 30 to 5, the accuracy declines dramatically. For 1v1 test, the best accuracy achieves 60% with dimension 30, while the worst accuracy is around 50% with dimension 5. For LOO test, the best accuracy achieves 80% while the worst accuracy is around 70%, which seems quite satisfying. Such kind of result suffices our imagination. But is that the best result?

4.2 P2P transfer

In P2P setting, we perform knowledge transfer from person i to person j respectively, which leads to 7 results per person. For simplicity, we calculate the mean accuracy of the 7 results. Besides, In order to validate the dimensionality reduction method, we set the features to be 5, 10, 15, 20 for the target dimension number of TCA. In comparison with the no transfer experiment, we also perform the 1v1 and



(a) P2P-1v1

(b) P2P-LOO

Figure 2: Experimental result of person to person transfer. (a) is the result of 1v1, the transfer result of person to person transfer, and (b) is the result of LOO, the transfer result of using 7 persons as source and the other person as target.

LOO version. Similar to no transfer version, 1v1 in this experiment means we use 1 person as the source domain while the other 7 (one by one) as the target version respectively, which leads to 7 results for one target domain. The accuracy for P1 means the average of 7 results where P1 is the target. In LOO version, we simply use 7 persons as the source domain and the other as the target. The result is as Figure 2(a) and Figure 2(b).

From the result, we can see that person to person transfer is of good accuracy (average accuracy is over 99.9%). This is mainly because the source and target domain are in the same feature space, proving that TCAdoes not lead to low performance when facing with same feature spaces. In addition, we tested the P2P performance against different dimension conditions (#dimensions after reduction = 5, 10, 20, 30), and TCA proves to have good dimensionality reduction benefits. Literally we can see that as the dimension changes from 30 to 5, the performances of every person does not change dramatically.

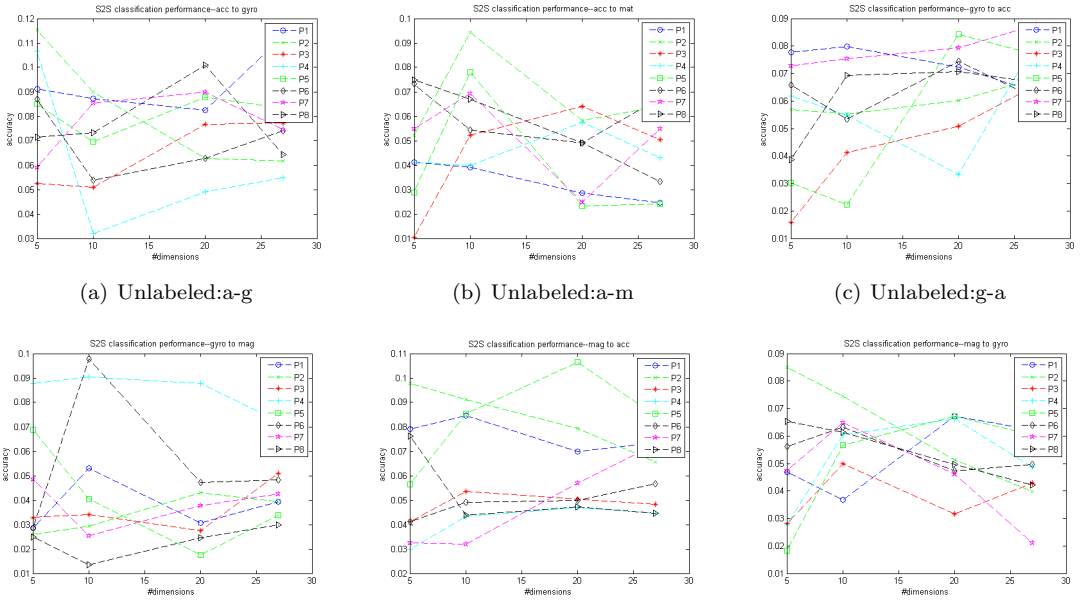
Another interesting point is that as the number of dimension increases from 5 to 30, we don't see any explicit increases of every person. Which means TCA based dimensionality reduction is very robust of dimensions.

When we do comparison between the results of no transfer and transfer (Figure 1(b) and Figure 2(b)), we do not see any considerable improvement. Their performances are almost the same, perhaps even better for no transfer version. For this phenomenon, we think maybe it's because TCA handles with different distribution of the data, and its target is to minimize the distance between source and target domain in the new feature representation, so when the source or target domain is composed of different distributions (in our case it's the source domain made of 7 different persons' data that causes the dramatically varied distribution), the new representation in new space will be not so fluent to obtain. And this leads to poor performance. However, the performance is still not so bad, proving that even with dramatically varied distributions, TCA can still preserve good results.

4.3 S2S transfer

In S2S setting, we perform knowledge transfer from accelerometer, gyroscope and magnetometer, respectively. For every round, we take 3-fold cross validation to calculate the transfer accuracy. There are 3 results for every person. The result tested on unlabeled data is as Figure 4.3 and the out-of-sample data is as Figure 4.3. Here we only test on torso part (there are other 4 parts remained untested).

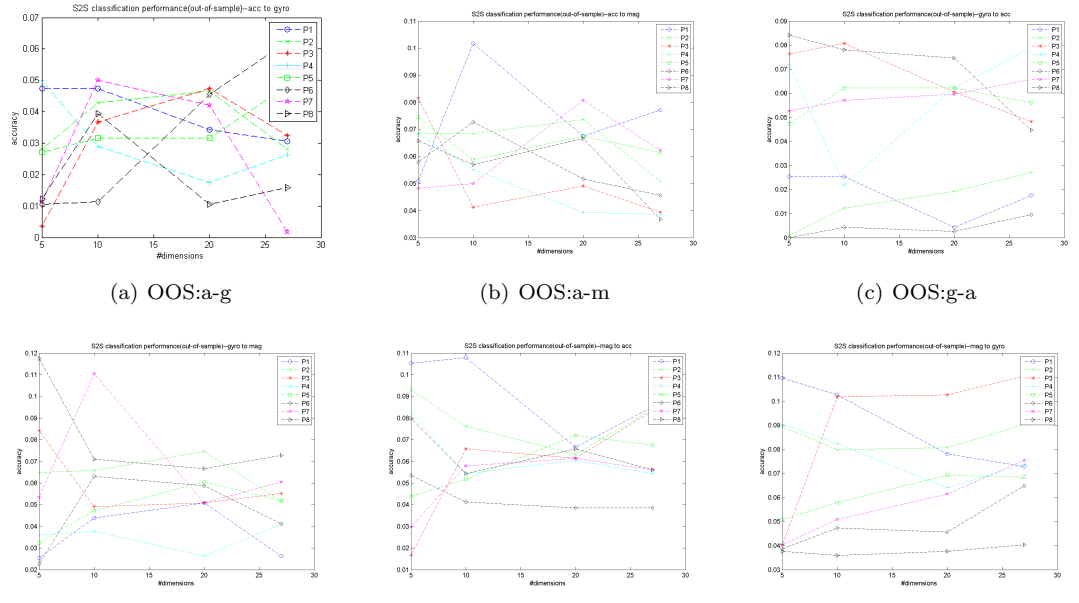
From the experimental results, we can see that TCA generates poor performance while transferring from sensor to sensor, with the average accuracy of below 10%. In fact, there are 19 class, random classification is still causing the accuracy of 5%! Such results mean that applying TCA directly to any two fields in different feature spaces does not yield good results unless the certain specified field can be analyzed according further knowledge. This throws light upon future research on transfer learning related



(a) Unlabeled:a-g (b) Unlabeled:a-m (c) Unlabeled:g-a

(d) Unlabeled:g-m (e) Unlabeled:m-a (f) Unlabeled:m-g

Figure 3: Test accuracy of S2S transfer on unlabeled data



(a) OOS:a-g (b) OOS:a-m (c) OOS:g-a

(d) OOS:g-m (e) OOS:m-a (f) OOS:m-g

Figure 4: Test accuracy of S2S transfer on out-of-sample data

activity recognition where researchers should pay much attention to the inner information of each field.

5 Conclusion

Our experiments show that TCA is able to perform knowledge transfer when source domain and target domain are in different feature distributions. The new feature space after TCA is in respectively low dimension, making it possible to perform dimensionality reduction while transferring knowledge. However, applying TCA directely on any other fields where source and target domain are in different feature spaces will generate poor performance, making it future work of doing more deep research on heterogeneous transfer learning.

Acknowledgment

Authors would like to thank Sinno Jialin from Nanyang Technological University for his supportive advice on transfer learning. Authors thank Wang Kong from Tsinghua University for his brilliant suggestion in mathematical problems, and Lisha Hu from ICT, CAS for her kind support of the experiments. And authors thank the teacher for her affectionate and inspiring class teaching.

References

[1] Diane Cook, Kyle D Feuz, and Narayanan C Krishnan. Transfer learning for activity recognition: A survey. *Knowledge and information systems*, 36(3):537–556, 2013.

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

[3] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.