ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Active transfer learning for activity recognition

Tom Diethe and Niall Twomey and Peter Flach

Intelligent Systems Laboratory, University of Bristol, UK

Abstract. We examine activity recognition from accelerometers, which provides at least two major challenges for machine learning. Firstly, the deployment context is likely to differ from the learning context. Secondly, accurate labelling of training data is time-consuming and error-prone. This calls for a combination of active and transfer learning. We derive a hierarchical Bayesian model that is a natural fit to such problems, and provide empirical validation on synthetic and publicly available datasets. The results show that by combining active and transfer learning, we can achieve faster learning with fewer labels on a target domain than by either alone.

1 Introduction

In this paper we are concerned with activity recognition, which is usually performed for the purposes of understanding the Activities of Daily Living (ADL) of a given individual. It is natural to consider the use of accelerometers for activity recognition, since it is clear that certain activities will have clear movement patterns for different parts of the body, whilst the sensors are relatively low-cost, low-power, and have wide user acceptance [1].

There are at least two major challenges for machine learning in this setting. Firstly, the deployment context will necessarily be very different to the the context in which learning occurs, due to individual differences patterns of motion for a given activity. Secondly, accurate labelling of training data is an extremely time-consuming process, and the resulting labels are potentially noisy and error-prone.

Our contributions are: 1) We provide a hierarchical Bayesian model for the transfer learning problem; 2) We combine this with myopic active learning; 3) Using this approach we show empirically that we can adapt to new domains using only a few labelled examples.

1.1 Related Work

Early work on active learning [2] demonstrated that it is possible to compute the statistically 'optimal' way to select training data, with the observation that the optimality criterion sharply decreases the number of training examples the learner needs in order to achieve good performance. This differs from the many heuristic methods for choosing training data, including choosing places where we don't have data, where we perform poorly, where we have low confidence *etc.*

Within a Bayesian framework, active learning can be naturally conceived since uncertainty is directly modelled, and there has been much interest in this area, particularly with respect to nonparametric methods. A major assumption in the majority of machine learning methods is that the training and deployment data are drawn from the same underlying distribution. For our application this assumption clearly does not hold, and so, knowledge transfer, if done successfully, would greatly improve the performance of learning by avoiding the costly acquirement of labels [3]. It is well known that the hierarchical Bayesian framework can be adapted to sequential decision problems [4], and it has been shown more recently that it provides a natural formalisation of transfer (reinforcement) learning [5].

Recently, [6] investigated active transfer learning for cross-system recommendation, since a newly launched system has a cold-start problem, where existing rating information is available. The authors construct entity correspondences with limited budget by using active learning to facilitate knowledge transfer across systems.

2 Hierarchical Bayesian Active Transfer Learning

We present here a multi-class extension of the Bayes Point Machine (BPM) [7], which is a Bayesian model for classification that makes the following assumptions: 1. The feature values $\mathbf{x}$ are always fully observed. 2. The order of instances does not matter. 3. The predictive distribution is a linear discriminant of the form $p(y_i|\mathbf{x}_i, \mathbf{w}) = p(y_i|s_i = \mathbf{w}'\mathbf{x}_i)$ where $\mathbf{w}$ are the weights and s_i is the score for instance i. 4. The scores are subject to additive Gaussian noise. 5. Each individual has a separate set of weights, drawn from a communal prior. For the purposes of activity recognition, assumption 2 may be problematic, since the data is sequential in nature. The strength of the temporal dependence in the sequence will determine how costly this approximation is, and this will in turn depend on how the data is preprocessed The factor graph for this model is illustrated in fig. 1, where $\mathcal{N}$ denotes a Gaussian density for a given mean μ and precision τ, and Γ denotes a Gamma density for given shape k and scale θ. The factor indicated by $\int$ is the 'arg-max' factor, which is like a probabilistic multi-class switch. The additive Gaussian noise from assumption 4 results in the variable $\tilde{s}$ This is a hierarchical multi-class extension of the Bayes point machine [7], where we have a plate around the individuals that are present in the training set (R), who form the "community". Online learning is performed using the standard assumed-density filtering method of [4].

To apply our learnt community weight posteriors to a new individual we can use the same model configured for a single individual (*i.e.* $R = 1$) with the priors over weight mean $\mu_\mathbf{w}$ and weight precision $\tau_\mathbf{w}$ replaced by the Gaussian and Gamma posteriors learnt from the individuals in the training set. This model is able to make predictions even when we have not seen any data for the new individual, and it is also possible to do online training as we receive labelled data for the individual. By doing so, we can smoothly evolve from making generic predictions that may apply to any individual to making personalised predictions specific to the new individual.

Given a set of potentially noisy training examples $\mathcal{S} = \{(\mathbf{x}_i, y_i)_{i=1}^m\}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, we wish to learn a general mapping $\mathcal{X} \rightarrow \mathcal{Y}$, and we can iteratively select a new input $\tilde{\mathbf{x}}$ and request a label $\tilde{y}$.

We use two base methods in order to do personalised active learning. The first is a simple uncertainty sampling method, where we select points using the marginal predictive distributions of points in the pool closest to chance levels (*e.g.* 0.5 for a binary classifier). As a sanity check, we also do "certainty" sampling, where we choose the points that the classifier is most confident about. Secondly, we extend the method outlined by [8]. We make a myopic assumption, where we only seek to label one data point at a time from a pool of potential examples and define a cost matrix $\mathbf{C} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$, $C_{i,j}$ denotes the risk associated with classifying a point i as j, and $C_{i,i} = 0,\ \forall i \in \mathcal{Y}$. The total cost on the community training set is

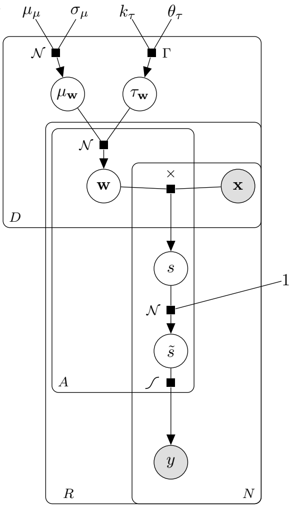


Fig. 1: Hierarchical community multi-class BPM.

$$J_S = \sum_{a\in\mathcal{Y}} \left(\sum_{i:y_i=a} C_{ai}(1-p_i) + \sum_{i:y_i\neq a} C_{ai}p_i \right)$$

The cost for an unlabelled point is

$$J_{\tilde{\mathbf{x}}_i} = \sum_{a\in\mathcal{Y}} \left(C_{ai}(1-p_i)p_i^* + C_{ia}p_i(1-p_i^*)\right) \approx \sum_{a\in\mathcal{Y}} \left((C_{ai}+C_{ia})(1-p_i)p_i\right),$$

where p_i^* is the *true* conditional density of the class label given the data point, which is approximated by p_i. The approximate misclassification cost is then $\frac{1}{m+1}(J_S + J_{\tilde{\mathbf{x}}_i})$. In the method of [9], the cost $L(\mathbf{x}_i)$ of acquiring a label for $\mathbf{x}_i$ is given a value in the same currency as the costs in $\mathbf{C}$. The expected value-of-information (VOI) criterion is then defined as

$$VOI(\tilde{\mathbf{x}}_i) = J_S + J_{\tilde{\mathbf{x}}_i} + L(\mathbf{x}_i) - L(\tilde{\mathbf{x}}_i).$$

Given a set of unlabelled points U, our strategy is to select cases for labelling and labelling method that have the highest VOI. Note that whenever $VOI(\mathbf{x}_{\hat{i}}) < 0$, we have a condition where knowing a single label does not reduce the total cost, which can be employed as a stopping criterion. We also evaluate the effect of considering the unlabelled example that yields the empirical greatest risk (VOI-) rather than the greatest relative risk (VOI+).

3 Experiments

Here we present experimental results that attempt to show that active learning can be especially useful in transfer learning settings, and whether the costs of the VOI method are justified. We first present some results on a synthetic dataset, and subsequently show the methods working the transfer between two publicly available activity recognition datasets from accelerometer data. Source code for all experiments is at https://github.com/IRC-SPHERE/ActiveTransfer.

3.1 Synthetic Experiment

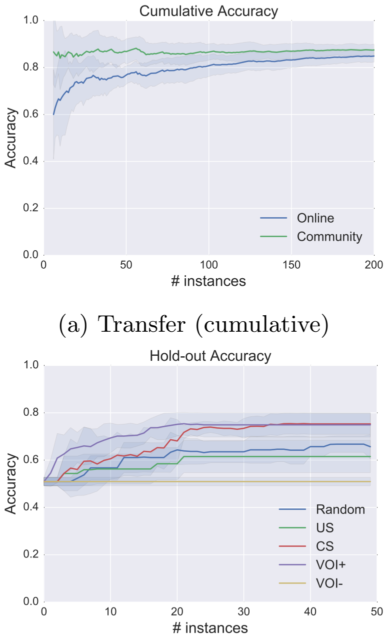
We sample from a community of 10 subjects using ancestral sampling as follows:

1. Create a community prior $\mathcal{N}(4,1)$, and sample a weight for each feature and subject from the community prior
2. For each subject and instance: a) Sample feature values uniformly in $[0,1]$, plus a constant bias (-1); b) Compute scores S as the inner product between features and weights; c) Compute the noisy score by sampling from $\mathcal{N}(S,1)$; and d) Threshold the noisy score at zero to compute the label.

The first 5 subjects are then used to train the hierarchical BPM shown in fig. 1. The remaining 5 are used in the online personalisation phase. For the weight priors we used $\mathcal{N}(0,1)$ means and $\Gamma(1,1)$ precisions. Inference is performed using Expectation Propagation (EP). To test the effectiveness of transfer learning using the hierarchical model, we compare using the posterior community weights as the priors for the personalisation phase with using the original $\mathcal{N}(0,1)$ priors, which will call `community` and `online` respectively.

In this synthetic dataset, every datapoint is equally useful for learning, since feature values are uniformly distributed. In real-world settings this is often not the case - there may either be high levels of redundancy between neighbouring examples (*i.e.* they are not truly independently and identically distributed), or some examples may simply be corrupted. In order to simulate this, we randomly set the feature vectors of 90% of the examples to zero for the dataset provided to the active learner.

Firstly we analyse the performance of the transfer learning method.In fig. 2a, we can see the cumulative online accuracy (and standard deviation (SD)) of a standard online method, versus using the community posteriors as the priors for the personalisation phase. In this setting we cycle through the test set, first predicting the label of an example, then observing its label an incorporating it into the model using an Assumed Density Filtering approximation [4].



(a) Transfer (cumulative)

(b) Active (hold-out)

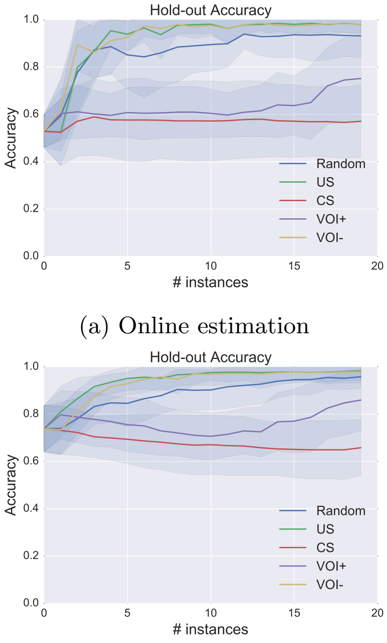
Fig. 2: Classification accuracy.

Secondly, we analyse the performance of the active learning methods in a myopic setting (see fig. 2b). The metric we use is the accuracy over the hold-out test set (200 examples), averaged over 5 subjects. Note that the VOI+ method learns fastest in this setting, whilst unsurprisingly the Certainty Sampling (CS) method outperforms the Uncertainty Sampling (US) method, since the US method will always choose the corrupted examples, whereas the CS method does the opposite.

Table 1: Publicly available data-sets for activity recognition based on body-worn accelerometers used in this study.

#	Ref.	Av. duration	Subjects	Type	Sampling rate	Labels
1	[10]	7 mins	30	Smartphone	50 Hz	Video
2	[11]	6 hours	14	MotionNode	100 Hz	Observer

3.2 Activity Recognition from Accelerometers



(a) Online estimation

(b) Active transfer

Fig. 3: Cumulative accuracy.

In transfer learning problems, it is common to refer to "source" and "target" data. The two datasets that will be used as source and target data are described below. Note that these were collected by different researchers, using different sensor equipment, and potentially labelled in different ways. As such, this is a good representation the challenge by researchers in this area when trying to apply models to new scenarios.

The source dataset [10], involved 30 participants aged 19-48 years and six activities were recorded. Each participant wore a smart-phone on the waist, with tri-axial linear acceleration and tri-axial angular velocity capture using its embedded accelerometer and gyroscope at a constant rate of 50 Hz. Annotation was done using video-recordings. The target dataset [11], involved 14 participants aged 21-49 years and 9 activities were recorded.

For both datasets, we considered the acceleration signals only. We extracted a total of 48 features based on the 'body' acceleration signal (see [10]. Posterior parameter distributions were first computed on the source dataset. These are then transferred to the target domain. It would not be rational to insist that the posteriors on the source and target domains are equally uncertain, and so the target precision distributions were set to the initial prior values ($\Gamma(1, 1)$) while keeping the source's posteriors over the weight means.

Fig. 3a shows the classification performance using online estimation with various instance sampling techniques (random, (un)certainty, VOI sampling). We can see that active parameter selection achieves greater classification performance and that it reaches its optimum at $\approx$ 5 samples. Furthermore, we can observe that the classification performance begins at 0.5. Fig. 3b presents the results obtained using posterior distributions from the source dataset. We first observe that the 'cold start' problem has been reduced as, without seeing examples from the new domain, active Bayesian transfer learning achieves approximately 70% accuracy (with equal class distributions). Also of interest is the variance of predictions, and we see that active selection methods consistently yield lower variance than random sampling. Over all experiments, active instance selection out-performs baseline methods. However, we note it is difficult to give a clear guide as to which active selection method should be chosen.

4 Conclusions

As we have seen, the activity recognition in the smart-home setting provides challenges in terms of the deployment context and accurate labelling of training data, which leads to a combination of active learning and transfer learning. We have argued that hierarchical Bayesian methods are particularly well suited to problems of this nature, and given a possible formulation of such a model. We have provided some experimental results on toy data do demonstrate the efficacy of the two components of this model. On real world data gathered using accelerometers, the more expensive VOI method failed to out-perform the simpler uncertainty sampling method, although for our purposes computational burden at the active learning stage is not a pressing issue. Our next steps will be to deploy the various active labelling methods in the prototype smart home, which will allow us to test the active learning framework, as well as the resident-to-resident transfer method. The house-to-house fusion and transfer on multi-modal sensor network can only be tested when multiple homes are available which will be the focus of future work.

References

[1] J.H.M. Bergmann and A.H. McGregor. Body-worn sensor design: what do patients and clinicians want? *Ann. of biomedical engineering*, 39(9):2299–2312, 2011.

[2] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *JAIR*, 4:129–145, 1996.

[3] S Pan. Transfer learning. In *Data Classification: Algorithms and Applications*, pages 537–570. 2014.

[4] M Opper. A Bayesian approach to on-line learning. pages 363–378. Cambridge University Press, New York, NY, USA, 1998.

[5] A Wilson, A Fern, and P Tadepalli. Transfer learning in sequential decision problems: A hierarchical Bayesian approach. In *ICML*, pages 217–227, 2012.

[6] Lili Zhao, Sinno Jialin Pan, Evan Wei Xiang, ErHeng Zhong, Zhongqi Lu, and Qiang Yang. Active transfer learning for cross-system recommendation. In *Proceedings of the* 27^{th} *AAAI Conference on Artificial Intelligence.*

[7] R Herbrich, T Graepel, and C Campbell. Bayes point machines. *JMLR*, 1:245–279, 2001.

[8] A Kapoor, E Horvitz, and S Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *IJCAI*, pages 877–882, 2007.

[9] P Rashidi and DJ Cook. Activity knowledge transfer in smart environments. *Pervasive Mob. Comput.*, 7(3):331–343, June 2011.

[10] D Anguita, A Ghio, L Oneto, X Parra, and JL Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *ESANN*, 2013.

[11] M. Zhang and A.A. Sawchuk. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *SAGAware*, 2012.