Bayesian Active Transfer Learning in Smart Homes

Tom Diethe TOM.DIETHE@BRISTOL.AC.UK
Niall Twomey NIALL.TWOMEY@BRISTOL.AC.UK
Peter Flach PETER.FLACH@BRISTOL.AC.UK
Intelligent Systems Laboratory, University of Bristol, UK

Abstract

There are at least two major challenges for machine learning in the smart-home setting. Firstly, the deployment context will be very different to the the context in which learning occurs, due to both individual differences in typical activity patterns and different house and sensor layouts. Secondly, accurate labelling of training data is an extremely time-consuming process, and the resulting labels are potentially noisy and error-prone. The resulting framework is therefore a combination of active and transfer learning. We argue that hierarchical Bayesian methods are particularly well suited to problems of this nature, and give a possible formulation of such a model.

1. Introduction and Motivation

One of the central hypotheses of a "smart home" is that a number of different sensor technologies may be combined to build accurate models of the Activities of Daily Living (ADL) of its residents. These models can then be used to make informed decisions relating to medical or healthcare issues. For example, such models could help by predicting falls, detecting strokes, analysing eating behaviour, tracking whether people are taking prescribed medication, or detecting periods of depression and anxiety. The Sensor Platform for HEalthcare in Residential Environment (SPHERE) project (`www.irc-sphere.ac.uk`) is developing a multi-modality sensor platform for smart homes with heterogeneous network connectivity. The SPHERE system uses three sensing technologies: environmental, video, and wearable devices. The data from each modality is collected in a gateway, which maintains time synchronisation in the system and controls data access to ensure user privacy. The current system is operational and is undergoing scripted validation experiments, where the sensor readings are processed to predict ADL against external (manual or automatic) activity labelling.

Proceedings of the 31st International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

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 both individual differences in typical ADL patterns, and also due to different house and sensor layouts. Secondly, accurate labelling of training data is an extremely time-consuming process (for example by manually annotating first person or third person video recordings), and the resulting labels are potentially noisy and error-prone. "Weaker" labelling can be achieved by requiring participants to self-report on the activities they are performing, either in real-time or *post-hoc*, but it may not be possible to verify the quality of such labels, and this is also potentially intrusive.

Multiple heterogeneous sensors in a smart-home environment introduce different sources of uncertainty, including failing sensors, biased readings, variable signal to noise ratio, etc. As a result we need to be able to handle quantities whose values are uncertain, and we need a principled framework for quantifying uncertainty which will allow us to build solutions in ways that can represent and process uncertain values. A compelling approach is to build a model of the data-generating process, which directly incorporates the noise models for each of the sensors. Probabilistic (Bayesian) graphical models, coupled with efficient inference algorithms, provide a principled and flexible modelling framework (Bishop, 2013).

2. Problem Definition

In this section we characterise the nature of the problems that arise in a smart-home environment, arguing that a combination of active and transfer learning is required.

2.1. Active Learning

Active learning is a paradigm of machine learning where the learner has control over the selection of training examples (or labels), rather than them being presented by nature (Cohn et al., 1996). An active learner may pose queries, usually in the form of unlabelled data instances to be labelled by an oracle (*e.g.* , a human annotator). Concretely, given a set of potentially noisy training examples

$\mathscr{S} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, m\}$, where $\mathbf{x}_i \in \mathscr{X}$ and $y_i \in \mathscr{Y}$, we wish to learn a general mapping $\mathscr{X} \rightarrow \mathscr{Y}$, and we can iteratively select a new input $\tilde{\mathbf{x}}$ (which may be from a constrained set) and request a label $\tilde{y}$. Active learning is well-motivated in many modern machine learning problems, where unlabelled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain, as is the case in the smart home setting.

Early work (Cohn et al., 1996) 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, where we expect it to change our model, and where we previously found data that resulted in learning (see (Cohn et al., 1996) for references). Note that this analysis gives statistical optimality for choosing the next $\tilde{\mathbf{x}}$ in terms of variance minimisation, but ignores the bias component, which can lead to significant errors when the learner's bias is non-negligible. Additionally, it doesn't allow for the inclusion of domain knowledge in any way.

Most active learning methods avoid model selection by training models of one type using one predefined set of hyper-parameters. An algorithm was proposed by (Ali et al., 2014) that actively samples data to simultaneously train a set of candidate models (different model types and/or different hyper-parameters) and also select the best model from this set. The algorithm actively samples points for training that are most likely to improve the accuracy of the more promising candidate models, and also samples points for model selection. This exposes a natural trade-off between focused active sampling that is most effective for training models, and unbiased sampling that is better for model selection. The authors empirically demonstrated on six test problems that this algorithm is nearly as effective as an active learning oracle with access to the optimal model.

2.2. Bayesian Active Learning

Active learning presents a scenario characterised by uncertainty: that is, we have uncertainty not only in training examples we have seen thus far, but also in the likely utility of different parts of the input space for improving our models. 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 such as Gaussian Processs (GPs). For example, in (Seo et al., 2000), a strategy of active data selection and test point rejection was used for GP Regression (GPR) based on the variance of the posterior over target values.

Information theoretic active learning has been widely studied for probabilistic models. For simple regression an optimal myopic policy is easily tractable (Krause & Guestrin, 2007), and central to this analysis was a theoretical bound which quantified the performance difference between active and *a-priori* design strategies. However, for other tasks and with more complex models, such as classification with nonparametric models, the optimal solution is harder to compute. Current approaches make approximations to achieve tractability. An approach that expresses information gain in terms of predictive entropy was applied to the GP Classifier (GPC) (Houlsby et al., 2011).

More recently, the problem of Bayesian active learning and experimental design was examined by (Javdani et al., 2014), where tests are selected sequentially to reduce uncertainty about a set of hypotheses. The authors argue that rather than minimising uncertainty, it is useful to consider a set of overlapping decision regions induced by these hypotheses, and the resulting goal is to drive uncertainty into a single decision region as quickly as possible.

2.3. Label Cost and Quality

Often we have "cheap" labels and "expensive" labels. In the SPHERE project, cheap labels come in two forms: residents will be given a smart-phone app through which requests can be made in real-time for labels by the system; or residents can be required to provide retrospective labels for there activities at certain times (*e.g.* at the end of the day). Expensive labels can be acquired by asking residents to wear a head-mounted camera for certain periods of time, and subsequently the video can be annotated by experts. Some work has been done on cost-sensitive active learning approaches that account for varying label costs while selecting queries. For example, (Kapoor et al., 2007) propose a decision-theoretic approach that takes into account both labelling costs and misclassification costs.

Cheap labels are often of poorer quality than expensive labels. Using the real-time method we may be able to get an instantaneous activity label, but we would have no information about the time course (*i.e.* start and end points) of the activity. Using the retrospective method we may get an approximate time course, but we are reliant on the memory and honesty of the individual. The issue of variable labelling quality was addressed by (Donmez et al., 2009), who modelled annotators as having different noise levels, and showed that both true instance labels and individual oracle qualities can be estimated. They then take advantage of these estimates by querying only the more reliable annotators in subsequent iterations active learning. However, this is not quite the same scenario as ours, since we have variable annotation methods rather than variable an-

notators.

2.4. Transfer Learning

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 the smart-home application this assumption clearly does not hold. In such cases, knowledge transfer, if done successfully, would greatly improve the performance of learning by avoiding the costly acquirement of labels. In recent years, transfer learning has emerged as a new learning framework to address this problem, and is related to areas such as domain adaptation, multi-task learning, sample selection bias, and covariate shift (Pan & Yang, 2010; Pan, 2014).

When learning and deploying models in a smart home environment, we have the problem that since getting labelled data is extremely expensive, we can only realistically get labelled data for a restricted set of homes and individuals. We then have two separate transfer learning problems:

1. Models learnt on a set of people to a new person
2. Models learnt on a set of houses to a new house

The two transfer learning problems have potentially different characteristics. For problem 1, different people will have different activity patterns, and will also likely perform certain activities in different ways. Furthermore, some activities will be much more prevalent for some individuals than others. For problem 2 different houses will have different house and sensor layouts, meaning that the order in which activities are performed will change, the durations of activities may change, and it will be extremely difficult to find a correspondence between sensors across the houses, even if they are in the same room. Problem 1 calls for learning about groups of individuals which we propose to solve using group-level hyper-priors which can then be transferred to a new individual.

Problem 2 can be tackled by manually introducing meta-features, and then the feature space is automatically mapped from the source domain to the target domain. In (Rashidi & Cook, 2011), the authors first assign a location label to each sensor indicating in which room or functional area the sensor is located. Then activity templates are constructed from the data for both the source and target data. Finally, a mapping is learnt between the source and target datasets based upon the similarity of activities and sensors. As an alternative, a recent study by (Feuz & Cook, 2014) introduced three heterogeneous transfer learning techniques that reverse the typical transfer model and map the target feature space to the source feature space. The authors evaluate the techniques on data from 18 different smart apartments located in an assisted-care facility and compares the results against several baselines, and argue that this method removes the need to rely on instance to instance or feature to feature co-occurrence data.

It is well known that the hierarchical Bayesian framework can be readily adapted to sequential decision problems (Opper, 1998), and it has also been shown more recently that it provides a natural formalisation of transfer learning (Wilson et al., 2012). The latter's results show that a hierarchical Bayesian Transfer framework significantly improves learning speed when tasks are hierarchically related within the domain of reinforcement learning. In another study (Gönen & Margolin, 2014), the authors formulated a kernelized Bayesian transfer learning framework that is a combination of kernel-based dimensionality reduction models with task-specific projection matrices, and aims to find a shared subspace and a coupled classification model for all of the tasks in this subspace.

3. Hierarchical Bayesian Active Transfer Learning

In this section we will develop a class of models that may be used to tackle the two types of transfer learning indicated above, and show how active learning might be performed using this model. These models draw on the insights of the studies presented in this paper, but are in themselves novel.

The multi-class Bayes Point Machine (BPM) (Herbrich et al., 2001) is a Bayesian model for classification, and 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 clearly sequential in nature. Intuitively, we might imagine that 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 (*i.e.* is raw data presented to the classifier, or are features instead computed from the time series?). It has been shown (Twomey et al., 2015) that under certain conditions structured models and unstructured models can yield equivalent predictive performance on sequential tasks, whilst unstructured models are also typically much cheaper to compute. The factor graph for this model is illustrated in Figure 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

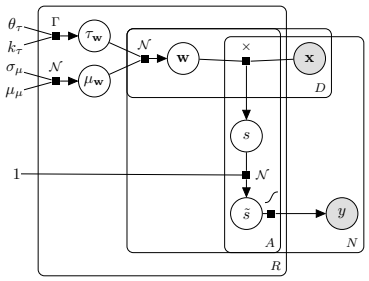

Figure 1. Hierarchical community-online multi-class Bayes point machine. A: Number of activities; R: Number of residents; N: Number of examples; D: Number of features.

and scale θ. The factor indicated by $\curlywedge$ is the 'arg-max' factor, which is like a probabilistic multi-class switch. The additive Gaussian noise from assumption 4 results in the variable $\tilde{s}$, which is a noisy version of the score s. This is a hierarchical multi-class extension of the Bayes point machine (Herbrich et al., 2001), where we have an extra 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 (Opper, 1998).

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, but 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. In training, the prior mean $\mu_{\mathbf{w}}$ is set to $\mathcal{N}(0,1)$ and the prior precision $\tau_{\mathbf{w}}$ is set to $\Gamma(4,0.5)$. The separate transfer learning problem from house-to-house is achieved through the method of introducing meta-features of (Rashidi & Cook, 2011), and then the feature space is automatically mapped from the source domain to the target domain. For simplicity, we have not shown this in Figure 1, and we assume that the features $\mathbf{x}$ are already these meta-features, and that for the personalisation phase the mapping has already taken place.

In order to do active learning to guide the choice of label acquisition in the online personalisation phase, we extend the method outlined by (Kapoor et al., 2007). Firstly we make a myopic assumption, where we only seek to label one data point at a time. Firstly we must define a cost matrix $\mathbf{C} \in \mathbb{R}^{A\times A}$, where A is the number of activities (classes), and $C_{i,j}$ denotes the risk associated with classifying a point i as j, and $C_{i,i} = 0, i = 1,\ldots,A$. The total cost on the community training set is

$$J_S = \sum_{a\in A}\left(\sum_{i:y_i=a} C_{ai}(1-p_i) + \sum_{i:y_i\neq a} C_{ia}p_i\right), \qquad (1)$$

where $p_i = p(\mathrm{sgn}(f(\mathbf{x}_i)) = 1|\mathbf{x}_i)$. The cost for an unlabelled point is

$$J_{\tilde{\mathbf{x}}_i} = \sum_{a\in A}\left(C_{ai}(1-p_i)p_i^* + C_{ia}p_i(1-p_i^*)\right), \qquad (2)$$

$$\approx \sum_{a\in A}\left((C_{ai}+C_{ia})(1-p_i)p_i\right), \qquad (3)$$

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 (Rashidi & Cook, 2011), 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 $\mathbf{C}$. Here we have n separate labelling methods with associated costs, which we will denote by $L_j(\mathbf{x}_i), j = 1,\ldots,n$. We must also define $G_j, j = 1\ldots,n$, where $0 \leq G_j \leq 1$, which quantifies the expected gain of a label given by labelling method j, where perfect labelling corresponds to $G_j = 1$. The expected value-of-information (VOI) criterion for a given labelling method j is then defined as

$$VOI(\tilde{\mathbf{x}}_i, j) = J_S + J_{\tilde{\mathbf{x}}_i} + \sum_{i=1}^{m} G_j L_j(\mathbf{x}_i) - G_j L_j(\tilde{\mathbf{x}}_i). \qquad (4)$$

Given a set of unlabelled points U, our strategy is to select cases for labelling and labelling method that have the highest VOI

$$(\hat{i}, \hat{j}) = \underset{i\in U, j\in\{1,\ldots,n\}}{\arg\max}\ VOI(\tilde{\mathbf{x}}_i, j). \qquad (5)$$

Note that whenever $VOI(\mathbf{x}_{\hat{i}}, \hat{j}) < 0$, we have a condition where knowing a single label does not reduce the total cost for a given labelling method, which can be employed as a stopping criterion if true for all methods simultaneously.

4. Conclusions

As we have seen, 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 observed that initial experiments on artificial data using this methodology give promising results. This is preliminary work: our next steps will be to deploy the various active labelling methods in the prototype SPHERE house, which will allow us to test the active learning framework, as well as the resident-to-resident transfer method. The house-to-house transfer method can only be tested when multiple homes are available, which will be in the latter stages of the SPHERE project.

References

Ali, Alnur, Caruana, Rich, and Kapoor, Ashish. Active learning with model selection. In Brodley & Stone (2014), pp. 1673–1679. ISBN 978-1-57735-661-5. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8547.

Bishop, C.M. Model-based machine learning. *Phil Trans R Soc A*, 371, 2013. URL http://dx.doi.org/10.1098/rsta.2012.0222.

Brodley, Carla E. and Stone, Peter (eds.). *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, 2014. AAAI Press. ISBN 978-1-57735-661-5. URL http://www.aaai.org/Library/AAAI/aaai14contents.php.

Cohn, David A., Ghahramani, Zoubin, and Jordan, Michael I. Active learning with statistical models. *J. Artif. Intell. Res. (JAIR)*, 4:129–145, 1996. doi: 10.1613/jair.295. URL http://dx.doi.org/10.1613/jair.295.

Donmez, Pinar, Carbonell, Jaime G., and Schneider, Jeff. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pp. 259–268, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557053. URL http://doi.acm.org/10.1145/1557019.1557053.

Feuz, Kyle Dillon and Cook, Diane J. Heterogeneous transfer learning for activity recognition using heuristic search techniques. *Int. J. Pervasive Computing and Communications*, 10(4):393–418, 2014. doi: 10.1108/IJPCC-03-2014-0020. URL http://dx.doi.org/10.1108/IJPCC-03-2014-0020.

Gönen, Mehmet and Margolin, Adam A. Kernelized Bayesian transfer learning. In Brodley & Stone (2014), pp. 1831–1839. ISBN 978-1-57735-661-5. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8132.

Herbrich, Ralf, Graepel, Thore, and Campbell, Colin. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, January 2001. URL http://research.microsoft.com/apps/pubs/default.aspx?id=65611.

Houlsby, Neil, Huszar, Ferenc, Ghahramani, Zoubin, and Lengyel, Máté. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011. URL http://arxiv.org/abs/1112.5745.

Javdani, Shervin, Chen, Yuxin, Karbasi, Amin, Krause, Andreas, Bagnell, Drew, and Srinivasa, Siddhartha S. Near optimal Bayesian active learning for decision making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pp. 430–438, 2014. URL http://jmlr.org/proceedings/papers/v33/javdani14.html.

Kapoor, Ashish, Horvitz, Eric, and Basu, Sumit. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pp. 877–882, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. URL http://dl.acm.org/citation.cfm?id=1625275.1625417.

Krause, Andreas and Guestrin, Carlos. Nonmyopic active learning of Gaussian processes: An exploration-exploitation approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 449–456, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273553. URL http://doi.acm.org/10.1145/1273496.1273553.

Opper, Manfred. On-line learning in neural networks. chapter A Bayesian Approach to On-line Learning, pp. 363–378. Cambridge University Press, New York, NY, USA, 1998. ISBN 0-521-65263-4. URL http://dl.acm.org/citation.cfm?id=304710.304756.

Pan, Sinno Jialin. Transfer learning. In Aggarwal, Charu C. (ed.), *Data Classification: Algorithms and Applications*, pp. 537–570. CRC Press, 2014. ISBN 978-1-4665-8674-1. URL http://www.crcnetbase.com/doi/abs/10.1201/b17320-22.

Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. ISSN 1041-4347. doi: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.191.

Rashidi, Parisa and Cook, Diane J. Activity knowledge transfer in smart environments. *Pervasive Mob. Comput.*, 7(3):331–343, June 2011. ISSN 1574-1192. doi: 10.1016/j.pmcj.2011.02.007. URL http://dx.doi.org/10.1016/j.pmcj.2011.02.007.

Seo, Sambu, Wallat, Marko, Graepel, Thore, and Obermayer, Klaus. Gaussian process regression: Active data selection and test point rejection. In *Mustererkennung 2000*, pp. 27–34. Springer, 2000.

Twomey, N., Diethe, T., and Flach, P.J. Understanding the role of structural modelling in sequence prediction. In *Mach. Learning and Knowledge Discovery in Databases - European Conf., ECML PKDD*, 2015.

Wilson, Aaron, Fern, Alan, and Tadepalli, Prasad. Transfer learning in sequential decision problems: A hierarchical Bayesian approach. In Guyon, Isabelle, Dror, Gideon, Lemaire, Vincent, Taylor, Graham W., and Silver, Daniel L. (eds.), *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, volume 27 of *JMLR Proceedings*, pp. 217–227. JMLR.org, 2012. URL http://jmlr.csail.mit.edu/proceedings/papers/v27/wilson12a.html.