An Interactive Learning Scenario for Real-time Environmental State Estimation Based on Heterogeneous and Dynamic Sensor Systems

Agnes Tegen, Paul Davidsson, and Jan A. Persson

Malmö University, School of Technology, Sweden
Internet of Things and People Research Center
{firstname.lastname}@mau.se

With the ongoing advances in the area of Internet of Things, the number of devices with sensors streaming data in our surroundings is growing rapidly. This will create new possibilities in continuously monitoring the state of the environment. However, this increasingly more complex setting is also posing new challenges, e.g. how to properly fuse data from different types of sensors with uncertain availability.

We are focusing on a setting where the task is to do real-time continuous estimations of certain aspects of the state of an environment. These estimations are based on data streams from a heterogeneous and dynamic set of sensors in that environment. Typically, data from different types of sensors needs to be fused in order estimate the aspect. For instance, within an office setting this could be what type of activity is currently taking place in a room or the number of people in a certain area of a building. In previous work [1], the concept of dynamic intelligent virtual sensors was suggested as a framework for data fusion. Common for many scenarios of this type, is that there is no available model that fuses the data and estimates the desired aspect of the state of the environment. Thus, such a model needs to be learned based on the streamed data provided by the sensors. Although the sensors may generate large amounts of data, there is typically a lack of labeled data that can be used for supervised learning.

The interactive learning challenge described above has been identified within an ongoing project with a number of industrial partners, partially validating its relevance to many real world applications.

1 Application Scenario Description

In a given environment, we define the set of all sensors that generates data as $S = \{s_1, s_2, \ldots\}$. While all sensors in S produce at least one instance of data, note that they do not necessarily have available data at all times. We also introduce a set $S_t \subseteq S$, containing all sensors from which data is available regarding the current point in time t. The data is generated by the sensors in a sequential fashion. Each instance of data contains the following information: *id* (unique identifier for the sensor), *data* (numerical or categorical measurement of the environment), *ts* (timestamp, the point in time when the data was measured according to the device).

We define a set of states, Y, representing the possible values of the aspect of the environment that we are interested in. If the entire state set is known beforehand, it can be defined along with the task. If not, the state set has to be defined over time, as labeled data becomes available. The labels provided by a user, denoted $y_{ts} \in Y$, may be used as training data and can be seen as ground truth for the state of the environment at time ts. They can be provided by the user's own initiative or when queried by the learner. The labeled data are stored for a certain time c_t, which can be set based on e.g. data storage possibilities.

The problem discussed in this paper can now be stated as follows: At any given point in time t, the task is to maximise the accuracy of the estimation of an aspect of the current state of the environment $\hat{y}_t \in Y$, using data from S_t, as well as labels and data, where $(t - c_t) \leq ts < t$, as input.

2 Discussion

In active learning the learner asks an oracle to label data [3]. Compared to the more common pool-based setting, where the learner starts with all unlabeled data and can ask for labels in any order, our setting is stream-based. Starting with a shortage, or non-existence, of labeled data, the algorithm must learn and adapt over time, as labeled data gradually becomes available. At each point in time the learner must decide whether to query or not, as it is not possible to query for old data points. Still, how much and when to query needs to be balanced, as there is a cost attached to it. Also, the aspect of reliability of the provided labels has to be considered, as human feedback is typically noisy. Different humans do not always agree on definitions or even with themselves over time.

Another complicating factor of the problem is that the entire state set might not be known from the beginning. This means that the learner must query not only to obtain training data, but also to learn the state space. If, for instance, the algorithm is not able to classify a state at a given time with sufficient certainty, it could query the user for the current state.

Generally, the different sensors do not provide data at synchronized points in time, as some are time triggered, with non-standardized time intervals, while others are event triggered. Furthermore, the timestamp attached to each data point is based on the sensor's individual clock. The learner, or a preprocessing step, needs to accommodate for this and align the data time-wise.

Sensor intensive systems can produce large amount of data and while it might be possible to store some data for a specified length of time, chances are all data cannot be stored indefinite. The learner must therefor incorporate the knowledge obtained from the data, so that it is not lost if the data is later discarded.

Transfer learning techniques is another way to handle the shortage of labeled data. For instance, if a model is trained to classify an aspect based on data from sensors in room A, the model could be adapted to do the corresponding task in room B. While transfer learning has been used successfully in many applications, the case where the input feature space for source and target (e.g. sensors in room A and B respectively) differ, is still a relatively new field of research [2, 4].

Acknowledgements This work was partially financed by the Knowledge Foundation through the Internet of Things and People research profile.

References

1. Mihailescu, R.C., Persson, J., Davidsson, P., Eklund, U.: Towards collaborative sensing using dynamic intelligent virtual sensors. In: International Symposium on Intelligent and Distributed Computing. pp. 217–226. Springer (2016)
2. Pan, S.J., Yang, Q., et al.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)
3. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
4. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data **3**(1), 9 (2016)