Multi-Instance Learning for Coarsely Labeled Time-Domain Inference

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

In the supervised learning setting, it is essential to have sufficient labeled data; however, in many domains, such as activity recognition, existing labeled data may not be available and the annotation process is often too cumbersome, time-consuming and prone to human error. In this work, we explore the use of Multiple-Instance Learning (MIL) in order to reduce the need for fine-grained labels. We examine the drop in performance on two existing time-domain gesture-annotated datasets and show that MIL given coarse-grain ground-truth annotations can achieve performance metrics comparable with standard supervised Machine Learning approaches given fine-grain labels. We evaluate the performance in a leave-one-participant-out fashion given (1) coarsely labeled field data, (2) finely labeled lab data and (3) coarsely labeled data from the held-out participant. Our analysis shows that we can achieve competitive performance given a small number of fine-grained labels in addition to many coarse-grained labels and that even very few labeled sessions from the held-out participant improve performance significantly. We use this to design a system that gives recommendations to developers on the granularity of the field data, based on an initial lab dataset.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

Multi-Instance Learning; Data Collection; Time-Domain; Activity Recognition; Eating Detection; Smoking Detection

INTRODUCTION

The ubiquity of mobile devices has led to a growing body of research in designing and solving gesture recognition tasks. These efforts have enormous implications in the mobile health community, self-tracking fitness industry and the development of state-of-the-art human-computer interfacing. The standard approach to gestural recognition employs an appropriate supervised classifier, which often performs exceptionally well given large amounts of labeled data and a well-chosen feature representation. The bottleneck to this approach is that acquiring sufficient gesture labels may be challenging, time-consuming or costly. While many techniques have been adopted to reduce the data annotation effort, this often comes at the expense of noisy labels due to factors such as human error.

A commonly used lightweight approach to gesture annotation is experience sampling [-1], where human subjects are prompted to label their current activity or recount their previous activity break-down. This is often best suited when the activities span a large enough time interval; otherwise, acquiring fine-grained labels remains difficult and especially prone to human error.

One of the most common solutions to reduce human error in data collection is video annotation. Although video labeling is relatively robust to human error, it is time-consuming, it introduces privacy concerns, and its power consumption is significantly large, making it impractical for collecting large-scale data in the field. Thus, there has been a significant effort to reduce the use of video recordings for annotated data collection while minimizing the label noise. Thomaz et al. [-1] employ an upward-facing camera mounted on a necklace to capture eating gestures in the field; the camera takes a snapshot of the subject every 30 seconds, significantly reducing the power consumption and labeling efforts required. Parate et al. [-1] use a 6-axis inertial sensor equipped on the upper arm in addition to a wrist-worn sensor in order to visualize the arm movements in a virtual 3D environment. This eliminates the need for video recordings while minimally increasing the risk of error. However, the annotation effort remains cumbersome and does not scale well to field data, because the additional armband is obtrusive.

Trabelsi et al. [-1] eliminate the need for training data altogether by using an unsupervised learning approach based on a Hidden Markov Model. While this technique achieves performance comparable to supervised learning approaches, it only provides a partition of the data by class and does not make precise label predictions in the absence of labeled data. When a large number of classes are present or positive labels are sparse, then sufficient annotated data once again becomes essential to realize robust, deployable classification systems.

Every submission will be assigned their own unique DOI string to be included here.

Recent work by Stikic and Schiele [-1] explores the feasibility of using Multi-Instance Learning (MIL) to reduce the labeling effort of activity recognition tasks while incurring minimal additional classification error. Although they show that comparable performance can be achieved with coarse-grained labels, they do not consider the case when the developers provide a small number of fine-grained labels in addition to field data.

In this work we demonstrate the on time-domain inertial data and evaluate the extent to which session-level and gesture-level labels improve performance. We additionally assess the boost in performance given a small number of fine-grained labels from the test user in a leave-one-participant-out evaluation. The

MULTI-INSTANCE LEARNING

In the Multi-Instance Learning (MIL) framework, we jointly consider instances, the atomic units over which predictions are made (i.e. gestures), and bags of instances, which may correspond to sessions or longer, manageable time intervals over which an activity is performed. In the binary setting, each bag is assigned a positive label if at least one instance in the bag is positive; bags with no positive instances are assumed to be negative.

The most naive MIL approach is Single-Instance Learning (SIL) [-1], which makes the usually false assumption that every instance in a positive bag is positive. This reduces the problem to a supervised instance-level classification task, which is generally done using a Support Vector Machine (SVM). When positive instances are sparse, the SIL assumption significantly hurts the classification performance.

In the activity recognition setting, Stikic and Schiele use the Maximum Pattern Margin Formulation (miSVM) originally proposed by Andrews et al. [-1] in order to account for the sparsity of positive bags. Due to the non-convexity of the objective function, they use a heuristic to learn the separating hyperplane. They initially train an SIL SVM, whose decision hyperplane is used to relabel the most positive predictions within positive bags. The SVM is then retrained on the relabeled data and the process is repeated until the labels converge. Although this approach accounts for the sparsity of positive gestures, it tends to over-predict the positive class [?] and has no mechanism to adjust the sensitivity based on known density.

Bunescu and Mooney [-1] deal with the challenge of sparse positive bags by using an adaptive SVM constraint (sMIL). In particular, they formulate the MIL constraint that there exists at least one positive instance in every positive bag X as follows

$$\begin{aligned} w\frac{\phi(X)}{|X|} + b \geq \frac{2-|X|}{|X|} - \xi_X \\ \xi_X \geq 0 \end{aligned}$$

where $w\frac{\phi(X)}{|X|} + b$ is the normalized prediction scores under the feature function ϕ, weights w and bias b, and ξ_X is the non-negative slack parameter that allows some extent of misclassification of instances in X to avoid over-fitting the model to the training data. When the bag size $|X|$ is small, the right-hand side becomes larger, suggesting that smaller positive bags are more informative.

Bunescu and Mooney additionally introduce a balancing parameter η, indicating the expected class distribution of instances within bags. The sparse balancing MIL (sbMIL) approach initially trains a sMIL classifier, then relabels the $\eta\,|X|$ most positive instances as positive and the remaining instances as negative. The final hyperplane is then learned using SIL given the relabeled data.

In this work we employ the sbMIL due to the sparsity of positive labels.

DATA

In order to reason in a practical sense about the trade-off between performance and labeling effort under the MIL formulation, we perform several evaluations on two existing datasets: the lab-20 eating dataset developed by Edison Thomaz [-1] and the RisQ smoking dataset developed by Parate et al. [-1]. In order to assess how well the model generalizes to unseen users, we perform leave-one-participant-out (LOPO) evaluations; that is, the model is trained on all but one participant and then evaluated on the held out participant.

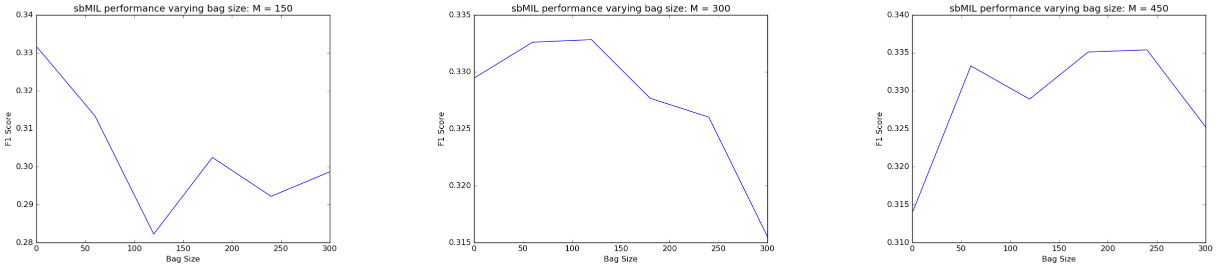
Lab-20 Eating

The lab-20 eating dataset comprises of 25Hz 3-axis accelerometer data collected using a wrist-worn inertial sensor from 20 individuals. Individuals were provided food to eat and were asked to perform other possible confounding actions as they please, including talking on the phone, brushing their teeth and combing their hair. The average duration across participants is 31 minutes 21 seconds and comprises of approximately 48% eating sessions. Note, however, that the proportion of eating gestures is much smaller, since non-eating gestures are frequently present within eating sessions.

We use Thomaz's evaluation as the baseline result for comparison. In his work, he uses a Random Forest classifier over 15 statistical features (mean, variance, skew, kurtosis and root mean square over each axis) extracted over windows of 6 seconds with 50% overlap. This generates 12379 labeled instances, of which 1480 (11.96%) are eating. He reports a 0.42 average f1 score over LOPO evaluations. We achieve similar performance using a linear SVM.

RisQ Dataset

The RisQ smoking dataset consists of 50Hz fused 9-axis inertial data in the form of quaternions from 14 subjects. The raw data stream is converted into a local trajectory in 3D space. Classification is done using a Random Forest, followed by a Conditional Random Field for smoothing predictions, over feature vectors of candidate windows identified by locating peak-trough-peak patterns indicative of smoking gestures. There are 11900 candidate windows, of which 358 (3.00%) are smoking gestures. A total of 37 features are extracted, including angular, velocity, displacement and duration features.



(a) M = 150

(b) M = 300

(c) M = 450

Figure 2: Average LOPO performance of sbMIL on Lab-20 dataset as a function of the bag size given 150, 300 and 450 additional labeled training instances respectively.

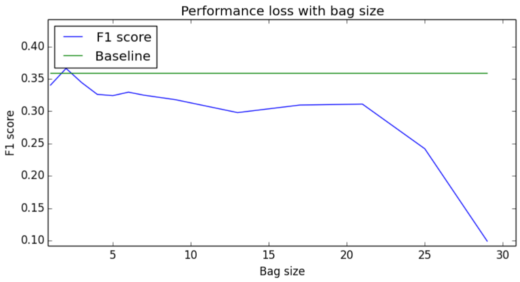


Figure 1: Average LOPO f1 score of sbMIL on Lab-20 dataset as a function of the bag size in blue; SVM f1 score in green.

Parate et al. report a LOPO precision of 91% and recall of 81%, which corresponds to f1 score of 85.7%.

In our work, we use the same pipeline but replace the Random Forest classifier with a sbMIL classifier to allow for sparse labels.

EXPERIMENTAL SETUP

In order to reason about the effectiveness of MIL techniques in gesture recognition, we evaluate the average LOPO performance for various bag sizes. Figure 1 shows that for the Lab-20 eating dataset as the bag size decreases, the performance of each MIL technique drops, and it is upper bounded by the baseline SVM performance.

Evidently, the performance is greater given more finely-grained labels. However, given that these labels may be difficult to acquire, we must ask: How many such labels do we need?

In order to address this, we evaluate several experiments in which M fine-grained labels are provided over a fixed number of participants and N coarse-grained labels are provided by the remaining participants. In the Lab-20 dataset, fine-grained labels are acquired from 5 participants and coarse-grained labels from the remaining 14 participants. In the RisQ dataset, they are acquired from 5 and 8 participants respectively. The coarse-grained labels may either be labeled sessions, which may vary in duration, or partitions of the data with a fixed duration. As a personalization step for enhancing performance, we additionally include K instances from the held-out participant in the training data, which are then excluded from the test set. Our experiments involve varying the values of N, M and K.

In each of the experiments, a subset of the training data is used and is therefore selected uniformly from the entire training data; to smooth out noise introduced by the randomness, the performance is averaged over 10 trials. The performance reported is in each case the best performance achieved using cross-validation over the model hyperparameters. These parameters include the expected class weights, the sparse balancing parameter η and the SVM regularization constant C.

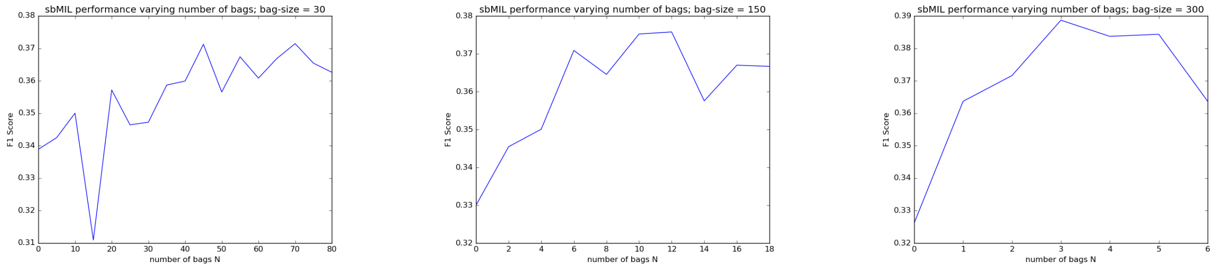
Note that the terms *instance* and *finely labeled data* are often used interchangeably, as are *session* and *coarsely labeled data*.

EVALUATION

Lab-20 Eating

Figure 1 shows in blue the average f1 score over all LOPO evaluations, varying the label granularity. The baseline standard SVM performance is shown in green for comparison. From Figure 1 it is clear that the performance drops very quickly as the granularity of the labels decreases.

However, Figure 2 demonstrates that this drop in performance is minimal even for large bag sizes, if in addition to coarse-grained labels, fine-grained labels are provided. More precisely, Figure 2 shows the average f1 score over all LOPO evaluations, varying the label granularity but fixing the number of finely labeled training instances M from the lab data. This is shown when 150, 300 or 450 labeled training instances are provided from the lab data. In each case, the number of training bags from the field remains constant but the granularity of the labels over those bags is varied. When $M = 150$, the f1 score drops noticeably; however, remains much larger than when no single instances are included, as shown in Figure 1. When $M = 300$, the performance drop is insignificant, and when $M = 450$, the performance remains roughly the same,



(a) bag size : 15 (1 min 30 s)

(b) bag size : 150 (15 min)

(c) bag size : 300 (30 min)

Figure 3: Average LOPO performance of sbMIL on Lab-20 dataset as a function of the number of bags given bag sizes of 15, 150 and 300 respectively and fixed number of labeled instances $M = 1500$

indicating that it may be acceptable to use field data with bag sizes of up to 300.

This alone could alternatively suggest that the additional field data we are providing does not give a significant boost in performance. To show that it indeed does increase the performance, we consider the case when the number of labeled training instances M is fixed and the number of training bags N varies.

Figure 3 shows the average LOPO f1 score on the Lab-20 eating dataset as the number of bags increases for bag sizes of 15, 150 and 300 instances. These correspond roughly to 1.5, 15 and 30 minute bags respectively. The number of labeled training instances is fixed at $M = 1500$. As the amount of training data increases, the f1 score increases, as expected. Interestingly, the performance is greater given larger bags, even when fewer labels are available. This suggests that many unlabeled instances are preferable to few labeled instances. This is the essential advantage of using MIL techniques.

Personalization

Lastly, we examine the effect of using a small number of instances and bags from the held-out user as training data. This enables us to personalize the model to the test individual. In a realistic setting, it would require that a new user label a small number of either instances or sessions. This can easily be done by performing the gesture while indicating it to the device in some way (i.e. voice command, holding a button while performing the gesture), or by indicating the start and end of a small number of sessions. In the former case, the user must be careful that the label is aligned with the gesture; otherwise, this will introduce noise and possibly hurt the performance of the classifier. In the latter case, the model is robust to human error and there is less burden placed on the user for personalizing the model.

RisQ Dataset

To show that this model generalizes well, we perform similar tests on the RisQ smoking dataset. In this case, we demonstrate that the performance increases as the number of labeled sessions increases. Figure 4 shows the average LOPO f1 score of the sbMIL classifier for various number of labeled sessions. We see nearly a 15% increase in performance given only 5 labeled sessions, comparable to the Random Forest/Condition Random Field baseline performance in the original RisQ pipeline.

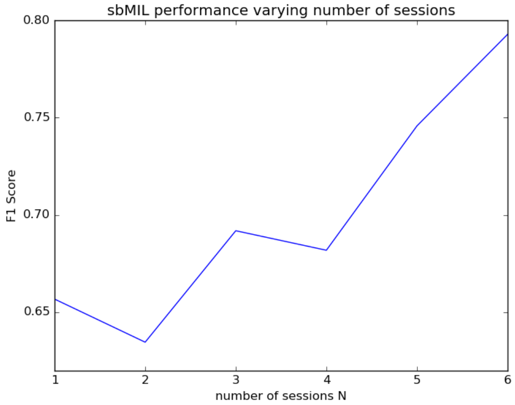


Figure 4: Average LOPO f1 score of sbMIL on RisQ dataset varying the number of labeled sessions

FUTURE WORK

In future work, we like to extend the lab-to-field system to address the issue of data transfer. In many cases the field data may diverge significantly from the lab data, in which case the label granularity recommendations may not be suitable.

We also plan to use this model in order to develop a large-scale 6-axis (accelerometer and gyroscope) activity dataset containing both finely labeled lab data and coarsely labeled field data. We are interested primarily in health-related gestures, including eating and drinking, exercise, washing and brushing teeth.

CONCLUSION

...

ACKNOWLEDGMENTS

...

REFERENCES

1. ACM. 1998. How to Classify Works Using ACM's Computing Classification System. (1998). http://www.acm.org/class/how_to_use.html.

2. R. E. Anderson. 1992. Social Impacts of Computing: Codes of Professional Ethics. *Social Science Computer Review December* 10, 4 (1992), 453–469. DOI: http://dx.doi.org/10.1177/089443939201000402

3. Anna Cavender, Shari Trewin, and Vicki Hanson. 2014. Accessible Writing Guide. (2014). http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/.

4. @_CHINOSAUR. 2014. "VENUE IS TOO COLD" #BINGO #CHI2014. Tweet. (1 May 2014). Retrieved Febuary 2, 2015 from https://twitter.com/_CHINOSAUR/status/461864317415989248.

5. Morton L. Heilig. 1962. Sensorama Simulator. U.S. Patent 3,050,870. (28 August 1962). Filed Februrary 22, 1962.

6. Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal and Ubiquitous Computing* 18, 4 (2014), 765–766. DOI: http://dx.doi.org/10.1007/s00779-014-0773-4

7. Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where Do Web Sites Come from?: Capturing and Interacting with Design History. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. ACM, New York, NY, USA, 1–8. DOI: http://dx.doi.org/10.1145/503376.503378

8. Nintendo R&D1 and Intelligent Systems. 1994. *Super Metroid*. Game [SNES]. (18 April 1994). Nintendo, Kyoto, Japan. Played August 2011.

9. Psy. 2012. Gangnam Style. Video. (15 July 2012). Retrieved August 22, 2014 from https://www.youtube.com/watch?v=9bZkp7q19f0.

10. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing*. ERIC, Bloomington, IN, USA.

11. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA.

12. Langdon Winner. 1999. *The Social Shaping of Technology* (2nd ed.). Open University Press, UK, Chapter Do artifacts have politics?, 28–40.