

LABEL DISAMBIGUATION AND SEQUENCE MODELING FOR IDENTIFYING HUMAN ACTIVITIES FROM WEARABLE PHYSIOLOGICAL SENSORS

Wei-Hao Lin and Alexander Hauptmann

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.

ABSTRACT

Wearable physiological sensors can provide a faithful record of a patient’s physiological states without constant attention of caregivers. A computer program that can infer human activities from physiological recordings will be an valuable tool for physicians. In this paper we investigate to what extent current machine learning algorithms can correctly identify human activities from physiological sensors. We further identify two challenges that developers need to address. The first problem is that the labels of training data are inevitably noisy due to difficulties of annotating thousands hours of data. The second problem lies in the continuous nature of human activities, which violates the independence assumption made by many learning algorithms. We approach the first problem of noisy labeling in the multiple-label framework, and develop a conditional Markov Models to take temporal context into consideration. We evaluate the proposed methods on 12,000 hours of the physiological recordings. The results show that Support Vector Machines are effective to identify human activities from physiological signals, and efforts of disambiguating noisy labels are worthwhile.

1. INTRODUCTION

With large amount of healthcare records in text, image, and video, multimedia technologies play an increasingly important role. For example, multimedia retrieval systems enable physicians to search patients’ records and medical information across multiple modalities [1]. Without automatic systems it will be extremely time-consuming and tedious to manually sift through huge collections of multimedia healthcare records.

Continuous recordings from wearable physiological sensors are of particular interest because they provide a long-term, close-body, and faithful records that few other modalities can offer. Hours of physiological signals can be obtained easily without disturbing patients and hiring extra caretakers.

Research prototypes [2, 3] and commercial products [4] have successfully shown the potential for monitoring physiological states of patients with wearable physiological sensors.

Physiological recordings, however, are of little use if they require huge human efforts to understand and interpret. In this paper we investigate the feasibility of automatically uncovering patients’ characteristics (e.g. gender, smoker) and identifying human activities (e.g. sleep, watching TV) from continuous physiological recordings. Our objective is to identify human activities that can be specified by physicians or patients; we thus approached the problems in a supervised learning framework, which is very different from previous work [5] that clusters physiological signals in an unsupervised fashion.

We identify two challenges posed by continuous physiological recordings in the tasks of identifying human activities. First, ambiguous and unannotated labels are abundant in real data. Instead of discarding data with noisy labels, we attempt to disambiguate noisy labels and incorporate them in the classifier learning process. Second, instead of simply treating every minute of physiological recording independent in time, which is definitely not true for most human activities, we build a conditional Markov model to exploit sequential relationship between physiological signals.

2. PHYSIOLOGICAL RECORDINGS

We evaluate our methods on the physiological recordings collected for the 2004 Physiological Data Modeling Contest (PDMC)¹. BodyMedia armbands, consisting of physiological sensors of acceleration, heat flux, Galvanic skin response, skin temperature, near-body temperature, are wore on the back of upper arms, and readings from each sensor are recorded every minute. Each physiological reading including nine numerical values from physiological sensors and two characteristics of the subject, resulting in 11-dimensional feature vector. The training set consists of 10,000 hours of recordings, and the testing set consists of 12,000 hours of recordings. Every minute of reading is manually annotated as unknown or

¹This material is based on work supported by the National Science Foundation (NSF) under Grant No. IIS-0121641.

¹<http://www.cs.utexas.edu/~sherstov/pdmc/>

one of the 51 activities, but only two activities, sleeping and watching TV, are officially evaluated on the testing set.

3. BASELINE SYSTEM

As a baseline, we approach the tasks of predicting patients’ characteristics or activities from physiological recordings as binary classification tasks. Each minute of physiological recordings is an input feature assumed to be independently drawn from a identical distribution. The data set consist of feature and label tuples, denoted as $\{(x_i, y_i)\}_{i=1}^n$, where x_i and y_i are the feature vectors and labels of the i -th example, and n is the size of data set. The labels are binary, for example, male or female, and presence or absence of a human activity of interest. Any classifiers can then be trained against the data set. In this paper we choose Support Vector Machines (SVM) [6], which has been shown to be very effective in a wide variety of classification tasks, including text classification [7] and image/video classification [8].

4. LABEL DISAMBIGUATION

One implicit assumption made by the baseline system in Section 3 is clean labels. Labeling physiological training data, however, is unlikely to be perfect. We distinguish two types of noisy labels: *ambiguous* labels and *unannotated* labels.

Ambiguous labels occur when a long session of recordings are annotated with a single label, but a short period within the session when an annotators does other activities are not marked. During a session labeled as “staying in the living room”, an annotator may temporarily watch TV but forget to annotate. If we are interested in building classifiers of “watching TV”, we should not treat all instances of “staying in the living room” as negative data. Since we cannot distinguish between labels that agree with true human activities and labels that do not agree, these labels are ambiguous.

Physiological data, especially continuous recordings from wearable physiological sensors, rarely are fully annotated. Labeling recordings minute by minute will create a huge cognitive load for annotators. 69.8% of our training data contain no labels. Annotators may also forget to annotation activities of interest, and classifiers will be penalized for ignoring positive and negative data with unannotated labels.

To disambiguate noisy labels we consider the following strategies:

All Equal We assign positive or negative labels with equal probability to instances with noisy labels, which is reasonable when no prior information is available.

All Negative We assign negative labels to all instances with noisy labels. Contrary to **All Equal**, we make a strong assumption that very few of noisy labels are positive. Treating unannotated labels as negative greatly increases

the number of negative training examples, which may improve classification accuracy.

Multiple Labels We make deliberate efforts to disambiguate noisy labels by considering how similar data with noisy labels to data with noiseless labels. Instead of making naïve assumptions in the previous two strategies, we treat noisy labels as *multiple labels*, that is, both positive and negative, and estimate how likely one is correct. The problem setup here is an instance of the multiple-label problems [9] with two labels. In the multiple-label framework, we optimize the Kullback-Leibler distance between the label conditional distribution, $\hat{p}(y|x_i)$, and the prediction from the model, $p(y|x_i, \theta)$, with parameters θ :

$$\theta^* = \arg \min_{\theta} \sum_i^n \sum_y \hat{p}(y|x_i) \log \frac{\hat{p}(y|x_i)}{p(y|x_i, \theta)}$$

Unlike supervised learning, $\hat{p}(y|x_i)$ is unknown and needs to be estimated, which leads to Expectation Maximization like algorithm. We initialize the label distribution randomly to train a first classifier. The learned classifier then updates the label distribution, and we re-train the classifier with new label distributions.

5. SEQUENCE MODELING

In addition to clean label assumption, the other drawback of the baseline system in Section 3 is the ignorance of sequential relationship between physiological signals. Human activities of interest, for example, sleeping and running, do not occur randomly. A user who enters a sleep state now will be more likely to stay in the same state for the next few minutes, and we should not assume the physiological signals to be independent temporally.

Inspired by McCallum et al. ’s work [10], we develop a conditional Markov model based on SVM to capture sequence relationship between physiological signals and states. Given a session of observations, i.e. feature vectors x_1, x_2, \dots, x_m , the tasks of predicting human activities can be formulated as finding the sequence of states, i.e. labels y_1, y_2, \dots, y_m , that maximizes the posterior probabilities,

$$\arg \max_{y_1, y_2, \dots, y_m} P(y_1, y_2, \dots, y_m | x_1, x_2, \dots, x_m; \theta) \quad (1)$$

where θ is the model parameters.

We make the Markov assumption that current state depends on only the previous state, but not earlier states. Contrary to Hidden Markov Models [11], we do not model the joint probability of states and observations. We let current states depend on both previous states and current observa-

tions. Under this model Eq. 1 can be rewritten as follows,

$$\arg \max_{y_1, y_2, \dots, y_m} P(y_1|x_1; \theta) \prod_{t=2}^m P(y_t|x_t, y_{t-1}; \lambda) \quad (2)$$

Conditional on y_{t-1} , we train conditional probability models $P(y_t|x_t, y_{t-1}; \theta)$ using SVM.

To find the most probable label sequence given a session of observations we use a Viterbi-like algorithm. Follow the notation in [11], denote $\delta_t(i)$ as the highest probability along a single path at time t and the state equals to q_i , where q_1 is positive, and q_2 is negative:

$$\delta_t(i) = \max_{y_1, y_2, \dots, y_{t-1}} P(y_1, y_2, \dots, y_t = q_i|x_1, x_2, \dots, x_t; \theta) \quad (3)$$

Eq. 3 can be efficiently solved using Dynamic Programming,

$$\delta_t(i) = \max_j \delta_{t-1}(j) \cdot P(y_t = q_i|y_{t-1}, x_t; \theta) \quad (4)$$

6. EXPERIMENTS

6.1. Baseline System

The baseline system in 3 is based on Support Vector Machines. Numerical values of feature vectors are scaled between zero and one. In the gender prediction task, all training data are fully labeled, and thus y_i are unambiguous. In human activity prediction tasks, SVM is trained against clean positive and negative data, and no ambiguous or unannotated data are used. Because gender does not change within a session, we take the majority vote from SVM’s predictions on each minute of the session. We use radial basis kernel for SVM, and grid searching on the held-out set is used to find the optimal values of two parameters (one for the kernel and one for cost).

The evaluation metric for the gender prediction task is balanced error rates; the evaluation metric for the activity identification tasks is weighted formula as specified by PDMC organizers². The random baseline is to guess the gender in every session as Gender 0, and to assign negative labels (majority label) for two activity identification tasks. We evaluate the baseline system on the training set in 10-fold cross-validation manner.

The results in Table 1 show that SVM is very effective for predicting gender and two activities, consistently outperform random baselines. Therefore physiological signals have great potential for monitoring and detecting the physical states of patients. Based on the degree of improvement over random baseline, the gender prediction task is much easier than two activity identification tasks, and “sleep” is easier to identify than “watching TV”. The classification accuracy appears to be positively correlated to the number of training examples and may explain the performance difference among three tasks.

²See <http://www.cs.utexas.edu/users/sherstov/pdmc/faq.html>

	Gender	Watching TV	Sleep
Random	0.5	0.7	0.7
SVM Baseline	0.9572	0.7548	0.8711
Improvement	+91%	+7.8%	+24.4%
Number of Training Sessions	1418	67	236

Table 1: The 10-fold cross-validation performance of the SVM baseline on the training set.

6.2. Label Disambiguation

We compare three disambiguation strategies in Section 4 on two activity identification tasks on both training and testing set. The label conditional probability for data with clean labels are pre-fixed, that is, either 1 or 0, and only noisy label distributions are updated. To prevent over-fitting we iterative until the classifier performance on the held-out set is not improved.

We implement the conditional label probabilities $\hat{p}(y|x_i)$ via sampling, that is, the label of each training example is sampled from the associated label probability. We obtain probability by fitting logistic regression on output values of a decision function of SVM [12]. The experimental results of two activity identification tasks are shown in Table 2.

		Watching TV	Sleep
Random		0.7	0.7
SVM Baseline	Training	0.7548	0.8711
All Equal	Training	0.7625	0.8834
	Testing	0.7314	0.9096
All Negative	Training	0.6957	0.8559
	Testing	0.7410	0.8999
Multiple Labels	Training	0.7613	0.8707
	Testing	0.7375	0.9125

Table 2: The performance of three label disambiguation strategies on the training and testing set.

First, “All Negative” are shown to be least effective label disambiguation strategy and worse than SVM baseline, partly because of the strong assumption made about noisy data. Multiple Labels and All Equal perform comparably, and consistently outperform SVM baseline on the testing set, which suggests that efforts spent on label disambiguation are worthwhile.

6.3. SVM-Based Markov Models

We implement a SVM-based Markov models in Section 5, and evaluate the models on the training set in a 10-fold cross-validation manner. The classification performance on both

gender and context tasks, however, is worse than or close to random baselines. The possible cause for low accuracy is because of highly unbalanced positive and negative examples after conditioning on s_{t-1} . For example, in Table 3, when

	$s_{t-1} = \text{neg}$	$s_{t-1} = \text{pos}$
$s_t = \text{neg}$	575776	75
$s_t = \text{pos}$	75	4338

Table 3: The number of examples for “watching TV” task in the training set.

conditioning on $s_{t-1} = \text{neg}$, we have seven thousands times more negative data than positive data. Similarly, we have fifty times more positive data than negative data when conditioning on $s_{t-1} = \text{pos}$. Unbalanced positive and negative data make estimation of the model $P(y_t = q_i | y_{t-1}, x_t; \theta)$ in Eq (4) very difficult.

7. CONCLUSIONS

In this paper we investigate the feasibility of monitoring and detecting human activities of patients from continuous physiological recordings using statistical learning algorithms. We identify two challenges posed by continuous physiological recordings: label ambiguity and sequential relationship, and propose three disambiguation strategies and SVM-based Markov models. The experiment results show that Support Vector Machines are very effective in both characteristic and activity identification tasks. By disambiguating noisy labels classification accuracy is further improved. Although sequential relationship is very strong for many human activities, unbalanced positive and negative examples makes learning very difficulty. We plan to experiment other sequence models, for example, Conditional Random Fields [13], to overcome the problem in future work.

8. REFERENCES

- [1] Kathleen R. McKeown, Shih-Fu Chang, James Cimino, Steven K. Feiner, Carol Friedman, Luis Gravano, Vasileios Hatzivassiloglou, Steven Johnson, Desmond A. Jordan, Judith L. Klavans, André Kushniruk, Vimla Patel, and Simone Teufel, “PERSIVAL, a system for personalized search and summarization over multimedia healthcare information,” in *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, 2001, pp. 331–340.
- [2] Eugene Shih, Vladimir Bychkovsky, Dorothy Curtis, and John Guttag, “Continuous medical monitoring using wireless microsensors,” in *Proceedings of the 2004 ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2004, p. 310.
- [3] Takuji Suzuki and Miwako Doi, “Lifeminder : An evidence-based wearable healthcare assistant,” in *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 2001, pp. 127–128.
- [4] Astro Teller and John (Ivo) Stivoric, “The bodymedia platform: Continuous body intelligence,” in *Proceedings of the First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, 2004.
- [5] Andreas Krause, Daniel P. Siewiorek, Asim Smailagic, and Jonny Farrington, “Unsupervised, dynamic identification of physiological and activity context in wearable computing,” in *Proceedings of the Seventh IEEE International Symposium on Wearable Computing*, 2003.
- [6] Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.
- [7] Thorsten Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in *Proceedings of the European Conference on Machine Learning (ECML)*, 1998.
- [8] Milind R. Naphade and John R. Smith, “On the detection of semantic concepts at TRECVID,” in *Proceedings of the Twelfth ACM International Conference on Multimedia*, 2004.
- [9] Rong Jin and Zoubin Ghahramani, “Learning with multiple labels,” in *Advances in Neural Information Processing Systems*, 2004, vol. 16.
- [10] Andrew McCallum, Dayne Freitag, and Fernando Pereira, “Maximum entropy markov models for information extraction and segmentation,” in *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2000.
- [11] Lawrence R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” in *Proceedings of the IEEE*, February 1989, vol. 77, pp. 257–286.
- [12] J.C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. 1999, pp. 61–74, MIT Press.
- [13] John Lafferty, Andrew McCallum, and Fernando Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 2001.