act2vec: Unsupervised Feature Learning for Human Activity Signals

Abstract

To better understand the role of physical activity and sleep in health requires the analysis of datasets which include both the time-series of longitudinal physical activity, and other clinical data. These datasets are not common as there is a gap between lifestyle data (e.g. sleep, physical activity) and clinical data normally captured in Electronic Health Records. This can be seen as a bottle neck for the development of machine learning approaches that link behavioral and clinical data. To overcome the problem of the unavailability of clinical data from a major fraction of subjects and unrepresentative subject populations, we propose an unsupervised time-series representational learning technique, `act2vec`. `act2vec` projects the time-series into a continuous vector space taking into account the co-occurrence of activity levels along with periodicity of human activity patterns at different levels of time granularity. Empirical evaluation shows that our proposed method performs and generalizes substantially better than the conventional time-series symbolic representational methods on multiple disorder prediction tasks.

Introduction

Physical activity and sleep are crucial to human wellbeing. The benefits of physical activity are paramount, including prevention of physical and mental disorders (Warburton, Nicol, and Bredin 2006). Many chronic conditions such as diabetes and schizophrenia, do include physical activity as a key self-management aspect to sustain and regain quality of life (Sigal et al. 2006). Sleep deprivation and poor sleep quality severely impact life quality (McClain et al. 2014).

In order to identify adverse health disorders, subjects have to go through a diagnosis phase. Medical practitioners have to rely heavily on a subject's recall of physical activity and sleep patterns ranging over several weeks to suggest diagnosis. With the increasing popularity of wearables we have an unprecedented ability to track a subject's physical activity and sleep patterns in real-time. The wearables market hits $14 billion in 2016 and is expected to rise to $34 billion by 2020 (Lamkin 2016). With over 411 million expected shipments in 2020, a huge proportion of population can be expected to possess wearables. Hence, we are on a cusp of a revolution with an ability for real-time monitoring by using the activity signals for various aspects of healthcare, including detection of serious conditions, their diagnosis, recommendations of therapy, and their monitoring, among others.

With long waiting times, often running into months just for the diagnosis stage, especially for sleep-related disorders, the economic cost is enormous (Shelgikar et al. 2014). It is estimated that a major sleep disorder, namely sleep apnea syndrome, alone causes a loss of $150 billion per year (Watson 2016). Current diagnosis approach involves conducting Polysomnography (PSG) during an overnight stay in a sleep lab, with multiple sensors attached to the subject's body. An automated tracking system that gets human activity data from wearable devices in real-time and mines the activity patterns can go a long way in timely diagnosis or risk evaluation of a subject, with huge potential savings towards health-care costs.

Clearly, this approach of using wearables for diagnosis of disorders has significant benefits. A major challenge, however, is that the diagnosis data is available only for a small fraction of the subjects who underwent the medical examination, mostly under a scare of disorder. Not only does it render useless the human activity data from majority of subjects, it also gives a very unrepresentative sample of disorder-positive population with respect to the general population. Hence, any approach towards using human activity signals should be designed to take into account the generalization of the approach. Task-specific supervised learning tends to generalize poorly with skewed datasets. The challenge is exacerbated by the noisy nature of activity signals and small dataset size of subjects who underwent diagnosis.

Taking the above challenges into account, we propose an unsupervised (task-agnostic) representational learning method `act2vec` for time-series data. `act2vec` uncovers the common patterns of human activity data by means of *distributed* representation, which can then be leveraged towards disorder diagnosis prediction tasks. The idea behind distributed representation learning is to transform the input data into an alternate space such that it can uncover the common patterns useful for classification and clustering tasks (Bengio, Courville, and Vincent 2013). It has shown enormous potential in areas like image classification, natural language processing, and speech processing (Hinton et al. 2012; Collobert et al. 2011; Krizhevsky, Sutskever, and Hinton 2012). As shown by these attempts in aforesaid do-

mains, adding unsupervised pretrained vectors to initialize the supervised models produces higher performance, with improved generalization and faster convergence. Distributed representation of time-series can capture semantic similarity between the time-series levels, and hence generalizes better.

In the domain of time-series analysis, a number of symbolic representation models exist, the most common being Symbolic Aggregate Approximation or SAX (Lin et al. 2007) and its variants. SAX converts each time-series window into a symbolic sequence by assuming a Gaussian distribution over the symbols. A number of time-series classification approaches based on deep learning have been proposed lately (Wang, Yan, and Oates 2016). Time-series vector-space models (*e.g.,* SAX-VSM) use TF-IDF (term frequency-inverse document frequency) vectors. To the best of our knowledge, our `act2vec` is the first work that uses a distributed representation learning approach to time-series.

One of the long standing challenges in the time-series domain is the selection of granularity for time-segments (also called time windows), which serve as the basic analysis units. We explore learning representations at various levels of time granularity, spanning over 30-seconds (device rate), an hour, a day, and a week. We devise a novel learning algorithm that optimizes three different measures to capture local and global patterns in a time-series along with a smoothing criteria. We use two publically available actigraphy datasets for training our model. We evaluate our approach against existing models and baselines on four disorder prediction tasks – Sleep Apnea, Diabetes, Hypertension, and Insomnia. Our experimental results show that our approach outperforms existing models by a good margin in all these tasks. Comparison between the variants of our models reveal that representation at the level of a day performs the best, and addition of smoothing criteria for a notion of periodicity in human activity helps improving the performance further.

In summary, our main contributions are: (*i*) we propose an unsupervised distributed representation learning framework for time-series analysis that beats existing time-series symbolic representation methods on multiple disorder prediction tasks; (*ii*) we investigate at different levels of time-granularity and show that day-level representations generalize best; and (*iii*) we devise an integrated learning algorithm that captures local and global properties of a time-series.

Related Work

We divide related works in three parts as described briefly below: (*i*) human activity recognition, (*ii*) representational learning, and (*iii*) time-series analysis methods.

Human activity research Human activity has been a widely studied area especially the problem of human activity recognition (HAR) with the goal of recognizing human activity from a stream of data such as camera recordings or sensors like motion detectors, gyro meters, and accelerometers. Wearable sensors like accelerometers (actigraphy) have mostly been used for human activity recognition task in machine learning (Bulling, Blanke, and Schiele 2014; Alsheikh et al. 2016), while medical practitioners use manual examination and extraction of features from the actigraphy data for diagnosing mostly sleep-disorders among subjects (Sadeh 2011). Recent work has tried using actigraphy data for quantifying sleep quality using deep learning (Sathyanarayana et al. 2016). The main difference with this method is that our approach is task-agnostic.

Representation Learning Bengio, Courville, and Vincent (2013) provide an overview of representation learning that is used to learn good features from the raw input space that are powerfully discriminative for downstream tasks. It is based on ideas of better network convergence by adding (unsupervised) pre-trained vectors and better encoding of mutual information of input features at the input layer (Goodfellow, Bengio, and Courville 2016). In past couple of years, the area has exploded with huge progress in natural language processing (Collobert et al. 2011), computer vision (Krizhevsky, Sutskever, and Hinton 2012), and speech recognition (Hinton et al. 2012). Of particular interest are the developments in natural language processing with distributed bag-of-words (DBOW) architectures (Mikolov et al. 2013) optimized to predict the context of the language unit (*e.g.,* word) at hand, unlike continuous-bag-of- words (CBOW) that predicts the language unit from its context. The DBOW model has been extended to incorporate discourse context (Saha, Joty, and Hasan 2017) and the node embeddings in networks (Grover and Leskovec 2016). We also use DBOW to capture local patterns in a time segment.

Time series analysis methods Time series methods use pair-wise similarity concept to perform classification and clustering tasks, with euclidean used as the measure of similarity. Dynamic Time Warping (Berndt and Clifford 1994) is a widely used technique for finding similarity between two time-series with totally different basal time units. However, it is extremely computationally expensive and its pair-wise similarity approach renders it non-scalable. This has lead to creation of time-series symbolic representation techniques like SAX (Symbolic Aggregate Approximation) (Lin et al. 2007), that convert time-series into a symbolic sequence that can be further used for feature extraction. SAX-VSM (SAX-vector space model) uses tf-idf (term frequency-inverse document frequency) transformation of these symbolic sequences to get vector representation of sequence windows. BOSS (Schäfer 2015) is a symbolic representation technique that uses Fourier transform of the time-series-windows to create symbolic sequences. BOSS-VS (Schäfer 2016) creates vector space model of BOSS sequence in a similar fashion to SAX-VSM. After a brief overview of related works, we next describe our modeling approach.

Our Approach

In order to create a representational schema for time-series activity signals, the first natural challenge we encounter is determining the right granularity of the analysis unit. For example, consider the time-series sample in Figure 1. Learning representations at the symbol level might result in sparse vectors that are too fine grained to be effective in the downstream tasks. Similarly, learning a single representation for the entire time sequence (*e.g.,* spanning a week) could result in generic vectors that lack the required discriminative

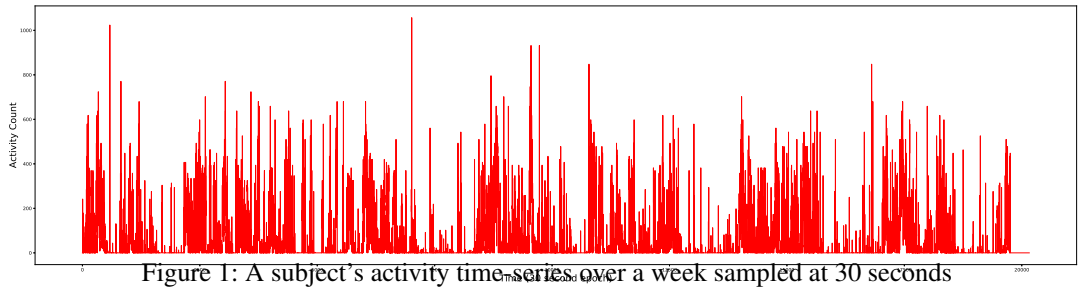

Figure 1: A subject's activity time-series over a week sampled at 30 seconds

power to solve the downstream tasks. As we will demonstrate later in our experiments that the right level of granularity is somewhat (*e.g.,* a day span) in between.

Considering units of analysis shorter than the sequence posits another challenge – how to capture the contextual dependencies in the representation. Since the units are parts of a sequence that describes a person's activity over a timespan (*e.g.,* week), they are likely to be interdependent. If such dependencies exist, the learning algorithm should capture this in the representation. In the following, we present our representation learning model that addresses these challenges.

Granularity of Time-Series Representation

As mentioned, for representing time-series data, it is important to consider the time-unit for which the embeddings are created. For example, for the activity signal, the granularity of analysis could be at the level of devices' sampling rate (30 seconds), an hour, a day, or a week. Each has its own advantages and disadvantages. Choosing analysis unit at the basal level (*i.e.,* device generated samples) can preserve fine grained information about the time series, but the resulting vectors could be sparse and susceptible to noise at the input source. On the other hand, a granularity at the week's scale would produce more condensed vectors, however at the cost of signal's low level information.

Let $\mathcal{D} = \{S_1, S_2, \cdots, S_P\}$ denote a time-series dataset containing activity sequences for P subjects, where each sequence $S_p = (t_1, t_2, \ldots, t_n)$ contains n activity measures for a subject p over a time period (a week in our case). Let $g \in \{30 \text{ seconds}, 1 \text{ hour}, 1 \text{ day}, 1 \text{ week}\}$ specify the granularity of the time span. We first break each sequence S_n into K consecutive time segments of equal length based on the value of g. Let $T_k = (t_a, t_{a+1}, \ldots, t_{a+L}) \in \mathcal{T}$ be such a segment of length L that starts at time a. Our aim is to learn a mapping function $\mathbf{\Phi} : \mathcal{T} \rightarrow \mathbb{R}^d$ to represent each time segment by a distributed vector representation of d dimensions. The vector representation for a full sequence can then be achieved by concatenating the K segment-level vectors.

In this study, we consider the following time spans along with the terminology followed for a comparative analysis:

- 30-second samples (`sample2vec`): This learns a distributed representation for each 30-second sample given by the device. Hence, our time-series of 20,160 length yields a representational space of $\mathbb{R}^{20160 \times d}$.
- Hour (`hour2vec`): It learns representation for the chunks of one-hour span of a time sequence. For a sequence of length 20,160, sampled at the rate of 30 seconds, `hour2vec` gives a vector space of $\mathbb{R}^{168 \times d}$.
- Day (`day2vec`): embeds time-series at the level of a day span, giving us a representational space of $\mathbb{R}^{7 \times d}$.
- Week (`week2vec`): provides embeddings at the scale of a week. A time series of length 20,160, sampled at the rate of 30 seconds, yields a vector in $\mathbb{R}^d$ space.

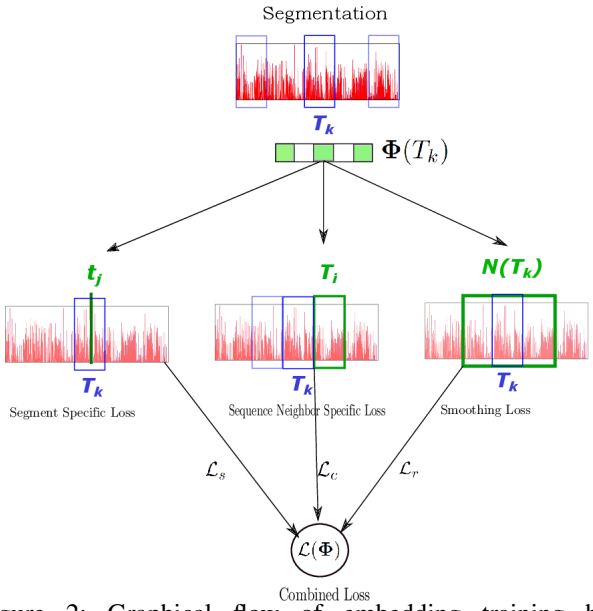

Figure 2: Graphical flow of embedding training by `act2vec`'s component objectives

For a given granularity level, we learn the mapping function $\mathbf{\Phi}$ by minimizing a loss that combines three components. Figure 2 presents the graphical flow of our model. In the following subsections, we first describe the component losses, and then we present the combined loss.

(1) Segment-Specific Loss

We use segment-specific loss to learn a representation for each time segment by predicting its own activity symbols. This is similar in spirit to the distributed bag-of-words (DBOW) approach of Le and Mikolov (2014). In this approach, activity symbols (analogous to 'words') and time sequences (analogous to 'sentences') are assigned unique

identifiers, each of which corresponds to a vector (to be learned) in a shared embedding matrix $\mathbf{\Phi}$. Given an input sequence $T_k = (t_a, t_{a+1}, \ldots, t_{a+L})$, we first map it to a unique vector $\mathbf{\Phi}(T_k)$ by looking up the corresponding vector in the shared embedding matrix $\mathbf{\Phi}$. We then use $\mathbf{\Phi}(T_k)$ to predict each symbol t_j sampled randomly from a window in T_k. To compute the prediction loss efficiently, Le and Mikolov use negative-sampling (Mikolov et al. 2013). Formally, the prediction loss with negative sampling is

$$\mathcal{L}_s(T_k, t_j) = -\log \sigma(\mathbf{w}_{t_j}^T \mathbf{\Phi}(T_k)) - \sum_{m=1}^{M} \mathbb{E}_{t_m \sim P(t)} \log \sigma(-\mathbf{w}_{t_m}^T \mathbf{\Phi}(T_k)) \quad (1)$$

where σ is the sigmoid function defined as $\sigma(x) = 1/(1 + e^{-x})$, $\mathbf{w}_{t_j}$ and $\mathbf{w}_{t_m}$ are the weight vectors associated with t_j and t_m symbols, respectively, and $P(t)$ is the noise distribution from which t_m is sampled. In our experiments, we use unigram distribution raised to the $3/4$ power as our noise distribution, in accordance to (Mikolov et al. 2013).

Since we ask the same segment-level vector to predict the symbols inside the segment, the model captures the overall pattern of a segment. Note that except for `sample2vec`, the model learns embeddings for both segments ('sentences') and symbols ('words'). With `sample2vec`, in the absence of any higher-level segment, the model boils down to the Skip-gram word2vec model (Mikolov et al. 2013) that learns embeddings for the symbols using a window-based approach. It is important to mention that segment-based approach is commonly used in time-series analysis, though among the representational models only vector space models like SAX-VSM (Senin and Malinchik 2013) and BOSS-VS (Schäfer 2016) look at the co-occurrence statistics at the segment level (indirectly), with a *bag-of-words* assumption.

(2) Sequence-Neighbor Specific Loss

The previous objective in Equation 1 captures local patterns in a segment. However, since the segments are contiguous and describe activities of the same person, they are likely to be related. For example, after a strenuous hour or day, there might be lighter activity periods. Therefore, representation learning algorithms should capture such relations between nearby segments in a time-series. We formulate this relation by asking the current segment vector $\mathbf{\Phi}(T_k)$ to predict its neighboring segments in the time-series: $\mathbf{\Phi}(T_{k-1})$ and $\mathbf{\Phi}(T_{k+1})$. Recall that each segment is assigned a unique identifier. If T_i is a neighbor to T_k, the neighbor prediction loss using negative sampling can be written as:

$$\mathcal{L}_c(T_k, T_i) = -\log \sigma(\mathbf{w}_{T_i}^T \mathbf{\Phi}(T_k)) - \sum_{m=1}^{M} \mathbb{E}_{T_m \sim P(T)} \log \sigma(-\mathbf{w}_{T_m}^T \mathbf{\Phi}(T_k)) \quad (2)$$

where, $\mathbf{w}_{T_i}$ and $\mathbf{w}_{T_m}$ are the weight vectors associated with T_i and T_m segments in the embedding matrix, respectively, and $P(T)$ is the unigram noise distribution over sequence IDs. As before, the noise distribution $P(T)$ for negative sampling is defined as a unigram distribution of sequences raised to the $3/4$ power.

(3) Smoothing Loss

While the previous two objectives attempt to capture local and global patterns in a time series, we also hypothesize that there is a smoothness pattern between neighboring segments. In some sense, it can also be viewed as a way to capture the periodicity of human activity. The learning algorithm should discourage any abrupt changes in the representation of nearby segments. We formulate this by minimizing the l_2-distance between the vectors. Formally, the smoothing loss for a time-segment T_k is

$$\mathcal{L}_r(T_k, \mathcal{N}(T_k)) = \frac{\eta}{\mid \mathcal{N}(T_k) \mid} \sum_{T_c \in \mathcal{N}(T_k)} \mid\mid \mathbf{\Phi}(T_k) - \mathbf{\Phi}(T_c) \mid\mid^2 \quad (3)$$

where, $\mathcal{N}(T_k)$ is the set of time-segments in proximity to T_k and η is the smoothing strength parameter. Note that the smoothing loss is not applicable to `week2vec`.

Combined Loss

We define our model as the combination of the losses described in Equations 1, 2, and 3:

$$\mathcal{L}(\mathbf{\Phi}) = \sum_{p=1}^{P} \sum_{T_k \in S_p} \sum_{\substack{t_j \in T_k \\ T_i \in \mathcal{N}(T_k)}} \Big[\mathcal{L}_s(T_k, t_j) + \mathcal{L}_c(T_k, T_i) + \mathcal{L}_r(T_k, \mathcal{N}(T_k))\Big] \quad (4)$$

We train the model using Stochastic Gradient Descent (SGD), iterating over the subjects in a dataset in each epoch. To estimate the representation of a segment, for each symbol sampled randomly from the segment, we take three successive gradient steps to account for the three loss components in Equation 4. By making the same number of gradient updates, the algorithm weights equally the contributions from the symbols in a segment and from the neighbors. Note that for `sample2vec` and `week2vec` only $\mathcal{L}_c$ loss is calculated since the other two objectives do not apply.

Experimental Settings

In this section, we describe our experimental settings – the prediction tasks on which we evaluate the learned vectors, the datasets, the models we compare and their settings.

Human activity time-series

The human activity data collected with a wearable devices (actigraphy) records mean activity count per base time-unit depending on the sampling rate of the device. The datasets we are working on, as described in next section, provide us with a signal that can only take integer values, unlike most time-series data, which makes embedding the input straightforward, without any pre-processing. Taking a leaf from speech recognition community, in case of floating point value signals, a preprocessing step of waveform extraction

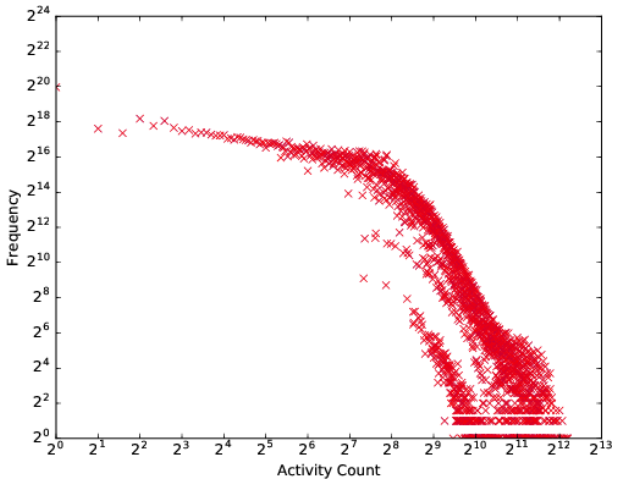

Figure 3: log-log plot of activity count and its frequency of occurrence in the time-series across our two data-sets

can be added. Actigraphy data is widely used for diagnosis of sleep disorders and quantification of physical activity for epidemiological studies.

Datasets

In this work, we use the Study of Latinos or SOL (Sorlie et al. 2010) and the Multi-Ethnic Study of Atherosclerosis or MESA (Bild et al. 2002) datasets. These datasets are made publicly available as a part of initiative to provide computer scientists with resources that can be used for helping the clinical experts (Dean 2nd et al. 2016). SOL has data for 1887 subjects ranging from physical activity to general diagnostic tests, while MESA has mostly the activity time-series data for 2237 subjects. National Sleep Research Resource provides these datasets at `sleepdata.org` with activity data (actigraphy) per subject for a minimum of 7 days measured with wrist-worn Philip's Actiwatch Spectrum device. Both the datasets contain time-series of activity counts for each subject sampled at a rate of 30 seconds.

While Figure 1 presents a sample activity signal, Figure 3 gives the log-log plot of activity signal value and its frequency in our combined datasets. A total of 1757 discrete values of signal were observed for our combined dataset. A very few missing values were observed in the dataset were replaced by unknown (UNK) token for training with our model. We only considered 7 days of data for each subject, since missing values increased enormously for subjects with more than 7 days of data.

Note: all the diagnosis prediction tasks were taken only from the SOL dataset, since MESA does not have the diagnosis data, public. We just use actigrpahy time-series from MESA for creating our embeddings.

Prediction Tasks

We evaluate the effectiveness of the learned embeddings on the following physical and mental disorder prediction tasks:

- **Sleep Apnea:** Sleep apnea syndrome is a sleep disorder characterized by reduced respiration during the sleep time, reducing oxygen flow to body. We use the Apnea Hypopnea Index (AHI) at 3% desaturation level. The patients having AHI$<$5 are characterized as *non-apneaic*, while those having an AHI $>$ 5 are characterized as *mild-to-severe-apnea* subjects. Sleep apnea detection is thus a two-class classification problem.
- **Diabetes:** Diabetes (type 2) is inability of body to respond to insulin, leading to elevated levels of blood sugar. Diabetes prediction is defined as a three-class classification problem, where the task is to decide whether a subject is a *non-diabetic*, *pre-diabetic*, or *diabetic*.
- **Hypertension:** Hypertension refers to abnormally high levels of blood pressure, an indicator of stress. Hypertension prediction characterizes a binary classification problem for increased blood pressure (BP). BP $>$ 140/90 is considered as having *hypertension*.
- **Insomnia:** Insomnia is a sleep disorder characterized by inability to fall sleep easily, leading to low energy levels during the day. We use a 3-class prediction problem for classifying subjects into *non-insomniac*, *pre-insomniac* and *insomniac* groups. We merged subjects suffering from moderate and severe insomnia into one class owing to very few subjects suffering from severe insomnia.

Models Compared

We compare our method with a number of naive baselines and existing systems that use symbolic representations:

(*i*) Majority Class This baseline always predicts the class that is most frequent in a dataset.

(*ii*) Random This baseline randomly picks a class label.

(*iii*) SAX VSM: Symbolic Aggregate Representation Vector Space or SAX-VSM (Senin and Malinchik 2013) combines SAX (Lin et al. 2007), one of the most widely used symbolic representation technique for time-series data with Vector Space Modeling using tf-idf (term frequency inverse document frequency) measure.

(*iv*) BOSS: Bag-of-Symbolic-Fourier-Approximation or BOSS (Schäfer 2015) is a symbolic representational learning technique that uses Discrete Fourier Transform (DFT) on sliding windows of time-series BOSS creates histograms of Fourier coefficients to create equal sized bins for the Fourier coefficients over the time-series, which are then assigned representational symbols. The classification method involves nearest neighbor approach, with labels assigned based on class that gets highest similarity score.

(*v*) BOSSVS: BOSS in Vector Space or BOSSVS (Schäfer 2016) is a vector space model similar to SAX-VSM except that it uses tf-idf vector space of the symbolic representation of the time-series obtained through BOSS.

BOSS is known to be one of the most accurate method on standard time-series classification tasks, with BOSS-VS performing marginally lower.

Variants of `act2vec`

We experiment with the following variants of our model:

Table 1: **Acc**uracy, **Pre**cision, **Rec**all, **Spec**ificity, and $\mathbf{F_1}$ values for **Sleep-Apnea** prediction for each method

Method	Clf.	Acc.	Pre.	Rec.	Spec.	F_1
Majority	0-R	74.6	00.0	00.0	100.0	00.0
Random		50.0	25.6	50.0	50.0	33.9
SAX-VSM		74.6	00.0	00.0	*100.0*	00.0
BOSS		70.4	30.0	12.5	90.1	17.6
BOSSVS		68.2	20.0	8.3	88.6	11.7
sample2vec	LR	50.0	27.8	54.0	48.5	36.7
hour2vec	LR	70.3	46.1	22.2	89.6	30.0
hour2vec+Reg	LR	71.4	36.8	14.3	91.4	20.5
day2vec	LR	61.9	32.8	42.0	69.1	36.8
day2vec+Reg	LR	65.1	39.6	38.2	76.1	**38.9**
day2vec+Reg	NB	61.3	38.3	*57.4*	62.9	**45.9**
week2vec	LR	*75.1*	*57.1*	8.3	97.9	14.5

***(i)* Unregularized models:** This group of models omit the smoothing component $\mathcal{L}_r(T_k, \mathcal{N}(T_k))$ in Equation 4. In the Results section, we refer to these models as sample2vec, hour2vec, day2vec, and week2vec.

***(ii)* Regularized models:** We perform smoothing in these models. This group includes hour2vec+Reg and day2vec+Reg. We omit sample2vec+Reg since it performed extremely poorly on all the tasks. Recall that smoothing is not applicable to week2vec.

Hyper-parameter selection

We have the following hyper-parameters: window size (w) for segment-specific (DBOW) loss, number of neighboring segments ($|\mathcal{N}(T_k)|$) and regularization strength (η) for day2vec and hour2vec. We tuned for $w \in \{8, 12, 20, 30, 50, 120, 500\}$, $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$, and $|\mathcal{N}(T_k)| \in \{2, 4\}$ on a development set containing 10% of the data. We chose w of size 20, 20, 30, and 50 for sample2vec, hour2vec, day2vec,and week2vec, respectively. The η of 0.25 and 0.5 were chosen for day2vec and hour2vec, respectively. The neighbor set size of 2 was selected for all the models. We selected an embedding size d=100 for all our models. In next section, we discuss our findings.

Results

In this section, we present the results of our representation learning models on the disorder prediction tasks. Since our goal is to evaluate the effectiveness of the learned vectors, we use simple linear classifiers to predict the class labels. Primarily, we use Logistic Regression (LR), which is a discriminative classifier. For our best representation learning model, we also present the results of a Naïve Bayes generative classifier since it showed considerable differences in performance with LR on our tasks. For the multi-class classification problems like Diabetes and Insomnia, we use One-vs-All classifiers tuning for micro-F_1 score. We ran each experiment 10 times and take the average of the evaluation measures to avoid any randomness in results.

Table 2: **Pre**cision (weighted), **Rec**all (weighted), F_1 values (weighted), and F_1-micro scores for the three class classification — non-diabetic, pre-diabetic, and diabetic — of **diabetes** prediction for each of time-series methods

Method	Clf.	Pre.	Rec.	F_1-macro	F_1-micro
Majority	0-R	23.7	*48.7*	21.7	31.9
Random		37.7	33.3	33.3	33.3
SAX-VSM		34.4	43.9	38.6	24.3
BOSS		39.1	38.8	38.9	31.5
BOSSVS		39.6	40.7	40.1	32.7
sample2vec	LR	41.2	38.9	40.0	36.7
hour2vec	LR	39.5	44.4	41.4	33.3
hour2vec+Reg	LR	40.8	43.9	*42.1*	32.0
day2vec	LR	41.2	40.7	40.9	38.0
day2vec+Reg	LR	44.7	40.7	41.8	**39.5**
day2vec+Reg	NB	43.5	39.2	40.6	36.5
week2vec	LR	40.8	44.4	40.6	34.1

Results for the four prediction tasks are shown in Tables 1, 2, 3, and 4. As we can observe across all the tasks, the day2vec+Reg outperforms all the models including the baseline time-series models. Across the board, models involving granularity on the scale of a day performs better than all the other granularities as well as baseline time-series methods. Clearly, among the act2vec variants, the week2vec models perform the worst, while hour2vec models perform just a bit better on an average. Hour- and week-level models perform around the same as the baseline time-series methods. The high-dimensional models based on samples (*i.e.,* sample2vec) perform better than hour-level, week-level, and baseline models. day2vec produces marginally better results than the sample2vec despite much lower dimensional space (2880x).

Intuition behind adding the smoothing loss to our model with Equation 3 was to test the hypothesis time-segments should be similar in structure representing a periodicity in human activities. As can be observed from the results, the regularization hypothesis was misguided at the sample- and hour-level segments. Using hour2vec+Reg leads to a significant drop (sleep apnea) or at par classification performance compared to hour2vec at its best. We omitted the sample2vec+Reg results for it only predicted the majority class on all the tasks. However, adding regularization helps produce gains across the board at the level of day2vec , our best act2vec model,

Another important aspect to note is the increase in generalization across classes on the prediction task of hour2vec and hour2vec+Reg. Our datasets are imbalanced with majority class being the subjects not suffering from the disorders under consideration, with classification task being to predict the disordering-positive subjects. Most of our models and baselines are highly biased towards predicting the majority class since data on these disorders were SOL's side-objective. hour2vec has lower accuracy but higher precision and recall than most of models, owing to its lower bias. Regularized hour2vec+Reg does better on F_1 scores while increasing the specificity/micro-F_1

Table 3: **Pre**cision (weighted), **Rec**all (weighted), F_1 values (weighted), and F_1-micro scores for the three class classification — no-insomnia, prethreshold-insomnia, and (moderate+severe) insomnia — of **insomnia** prediction for each of time-series methods

Method	Clf.	Pre.	Rec.	F_1-macro	F_1-micro
`Majority`	0-R	38.3	61.9	47.4	25.5
`Random`		46.6	33.3	33.3	33.3
`SAX-VSM`		38.3	61.9	47.4	25.5
`BOSS`		47.6	52.2	*49.8*	34.9
`BOSSVS`		45.2	50.1	47.5	33.1
`sample2vec`	LR	41.6	43.9	42.4	35.3
`hour2vec`	LR	42.5	52.4	44.6	28.5
`hour2vec+Reg`	LR	39.8	51.3	43.5	28.7
`day2vec`	LR	46.2	44.4	45.2	35.8
`day2vec+Reg`	LR	47.9	45.5	46.6	**39.7**
`day2vec+Reg`	NB	48.2	39.2	40.8	33.3
`week2vec`	LR	*51.5*	*55.0*	44.2	31.5

scores along with accuracy on all the prediction tasks than `hour2vec`, thus making it more generalized.

Clearly, the level of granularity makes a lot of difference to the performance of our models. From the above results on four different tasks we can conclude that while low granularity level (*e.g.,* `sample2vec`) suffered from coarse embeddings that got entangled at local level statistics while the high granularity (`week2vec`) level produced embeddings that lost the ability to discriminate. Another lesson being the smoothing at sample and hour levels. Regularization at sample level made them lose all the discriminative power for classification, which is understandable, considering the noise in activity levels at such a fine granularity. At the hour level too the assumption of nearby hours holding similar level of activity is flawed. One can expect such a pattern at the day level, since human activities tend to be similar across the consecutive days, as shown by improved performance of regularized `day2vec`.

Table 4: **Acc**uracy, **Pre**cision, **Rec**all, **Spec**ificity, and $\mathbf{F_1}$ values for **Hypertension** prediction for each method

Method	Clf.	Acc.	Pre.	Rec.	Spec.	F_1
`Majority`	0-R	74.9	00.0	00.0	100.0	00.0
`Random`		50.0	25.1	50.0	50.0	33.4
`SAX-VSM`		74.9	0.00	0.00	*100.0*	0.00
`BOSS`		*69.9*	35.2	25.5	84.5	29.6
`BOSSVS`		69.9	36.1	27.7	83.8	31.3
`sample2vec`	LR	51.3	33.3	48.3	52.7	39.5
`hour2vec`	LR	68.2	36.7	18.4	87.4	24.4
`hour2vec+Reg`	LR	68.2	36.0	17.0	88.2	23.1
`day2vec`	LR	60.8	39.1	41.7	69.8	40.3
`day2vec+Reg`	LR	68.2	*41.8*	45.0	76.8	**43.4**
`day2vec+Reg`	NB	59.3	38.0	*56.1*	60.6	**45.3**
`week2vec`	LR	67.7	58.3	11.1	96.0	18.7

Conclusions

Given the enormous popularity of wearable devices of human activity tracking, we are looking at a huge potential for personalized automated health-care that can not only reduce health-care costs but also help subjects avoid long waiting times. Such a system can potentially alert subjects with a risk of potential disorder, that can help in early treatment. Owing to absence of diagnosis data, majority of valuable activity data becomes ineffectual. Disorder detection also involves huge generalization issues like diagnosed subjects' population skew individual, and ethnic differences. In such a scenario, an unsupervised representational learning approach can effectively encode common human activity patterns in comparison to the end-to-end supervised learning approaches that may not generalize well.

We model human activity time-series data using a representational learning approach that can encode time-series at local (symbol/measurement) and global (*e.g.,* day) level. By testing our models at different levels of granularity on prediction task for commonly occurring disorders, we find that day level granularity preserves the best representations. Our model, the first task-agnostic representational learning time-series model using simple linear classifier, beats other symbolic representation models on the disorder prediction tasks. These time-series models are computationally expensive and hard to scale unless an expert feature extraction is performed, while our model learns the representational features automatically, giving better performance on our tasks using simple linear classifiers.

Future Work We plan to use alternative models of embeddings like convolutional neural network or auto-encoder based models. We also plan to use our embeddings to initialize the end-to-end deep learning models and make a comparison with a purely end-to-end based approach.

References

[Alsheikh et al. 2016] Alsheikh, M. A.; Selim, A.; Niyato, D.; Doyle, L.; Lin, S.; and Tan, H.-P. 2016. Deep activity recognition models with triaxial accelerometers. In *AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*.

[Bengio, Courville, and Vincent 2013] Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.

[Berndt and Clifford 1994] Berndt, D. J., and Clifford, J. 1994. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, 359–370. Seattle, WA.

[Bild et al. 2002] Bild, D. E.; Bluemke, D. A.; Burke, G. L.; Detrano, R.; Diez Roux, A. V.; Folsom, A. R.; Greenland, P.; JacobsJr, D. R.; Kronmal, R.; Liu, K.; et al. 2002. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology* 156(9):871–881.

[Bulling, Blanke, and Schiele 2014] Bulling, A.; Blanke, U.; and Schiele, B. 2014. A tutorial on human activity recog-

nition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46(3):33.

[Collobert et al. 2011] Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

[Dean 2nd et al. 2016] Dean 2nd, D.; Goldberger, A. L.; Mueller, R.; Kim, M.; Rueschman, M.; Mobley, D.; Sahoo, S. S.; Jayapandian, C. P.; Cui, L.; Morrical, M. G.; et al. 2016. Scaling up scientific discovery in sleep medicine: The national sleep research resource. *Sleep* 39(5):1151–1164.

[Goodfellow, Bengio, and Courville 2016] Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Grover and Leskovec 2016] Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864.

[Hinton et al. 2012] Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29(6):82–97.

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

[Lamkin 2016] Lamkin, P. 2016. Wearable tech market to be worth $34 billion by 2020. `https://www.forbes.com/sites/paullamkin/2016/02/17/wearable-tech-market-to-be-worth-34-billion-by-2020/`. [Online; accessed 17-July-2017].

[Le and Mikolov 2014] Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1188–1196.

[Lin et al. 2007] Lin, J.; Keogh, E.; Wei, L.; and Lonardi, S. 2007. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery* 15(2):107–144.

[McClain et al. 2014] McClain, J. J.; Lewin, D. S.; Laposky, A. D.; Kahle, L.; and Berrigan, D. 2014. Associations between physical activity, sedentary time, sleep duration and daytime sleepiness in us adults. *Preventive medicine* 66:68–73.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

[Sadeh 2011] Sadeh, A. 2011. The role and validity of actigraphy in sleep medicine: an update. *Sleep medicine reviews* 15(4):259–267.

[Saha, Joty, and Hasan 2017] Saha, T.; Joty, S.; and Hasan, M. 2017. Con-s2v: A generic framework for incorporating extra-sentential context into sen2vec. In *Proceedings of The European Conference on Machine Learning & Principles and Practice of knowledge discovery in databases*, ECML-PKDD'17. Macedonia, Skopje: Springer.

[Sathyanarayana et al. 2016] Sathyanarayana, A.; Joty, S.; Fernandez-Luque, L.; Ofli, F.; Srivastava, J.; Elmagarmid, A.; Taheri, S.; and Arora, T. 2016. Impact of physical activity on sleep: A deep learning based exploration. *arXiv preprint arXiv:1607.07034*.

[Schäfer 2015] Schäfer, P. 2015. The boss is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery* 29(6):1505–1530.

[Schäfer 2016] Schäfer, P. 2016. Scalable time series classification. *Data Mining and Knowledge Discovery* 30(5):1273–1298.

[Senin and Malinchik 2013] Senin, P., and Malinchik, S. 2013. Sax-vsm: Interpretable time series classification using sax and vector space model. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 1175–1180. IEEE.

[Shelgikar et al. 2014] Shelgikar, A. V.; Durmer, J. S.; Joynt, K. E.; Olson, E. J.; Riney, H.; and Valentine, P. 2014. Multidisciplinary sleep centers: strategies to improve care of sleep disorders patients. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine* 10(6):693.

[Sigal et al. 2006] Sigal, R. J.; Kenny, G. P.; Wasserman, D. H.; Castaneda-Sceppa, C.; and White, R. D. 2006. Physical activity/exercise and type 2 diabetes. *Diabetes care* 29(6):1433–1438.

[Sorlie et al. 2010] Sorlie, P. D.; Avilés-Santa, L. M.; Wassertheil-Smoller, S.; Kaplan, R. C.; Daviglus, M. L.; Giachello, A. L.; Schneiderman, N.; Raij, L.; Talavera, G.; Allison, M.; et al. 2010. Design and implementation of the hispanic community health study/study of latinos. *Annals of epidemiology* 20(8):629–641.

[Wang, Yan, and Oates 2016] Wang, Z.; Yan, W.; and Oates, T. 2016. Time series classification from scratch with deep neural networks: A strong baseline. *CoRR* abs/1611.06455.

[Warburton, Nicol, and Bredin 2006] Warburton, D. E.; Nicol, C. W.; and Bredin, S. S. 2006. Health benefits of physical activity: the evidence. *Canadian medical association journal* 174(6):801–809.

[Watson 2016] Watson, N. F. 2016. Health care savings: the economic value of diagnostic and therapeutic care for obstructive sleep apnea. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine* 12(8):1075.