ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/303790988

Exploring and Comparing Machine Learning Approaches for Predicting Mood Over Time

Conference Paper *in* Smart Innovation · June 2016

DOI: 10.1007/978-3-319-39687-3_4

CITATIONS
14

READS
1,003

6 authors, including:

Johnno Pastor
Vrije Universiteit Amsterdam
1 PUBLICATION **14** CITATIONS
SEE PROFILE

Mark Hoogendoorn
Vrije Universiteit Amsterdam
207 PUBLICATIONS **1,762** CITATIONS
SEE PROFILE

Jeroen Ruwaard
GGZ inGeest
66 PUBLICATIONS **1,653** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Reinforcement Learning in Regulated Domains View project

Project Interapy View project

All content following this page was uploaded by Joost Asselbergs on 07 November 2017.

The user has requested enhancement of the downloaded file.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/301296987

Exploring and Comparing Machine Learning Approaches for Predicting Mood over Time

Conference Paper · June 2016

READS

7

6 authors, including:

Jeroen Ruwaard

VU University Amsterdam

33 PUBLICATIONS **325** CITATIONS

SEE PROFILE

Heleen Riper

VU University Amsterdam

150 PUBLICATIONS **2,832** CITATIONS

SEE PROFILE

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.

Available from: Ward van Breda

Retrieved on: 25 April 2016

Exploring and Comparing Machine Learning Approaches for Predicting Mood over Time

Ward van Breda[1], Johnno Pastor[1] Mark Hoogendoorn[1], Jeroen Ruwaard[2],
Joost Asselbergs[2], and Heleen Riper[2]

[1] VU University Amsterdam, Department of Computer Science, De Boelelaan 1081,
1081 HV Amsterdam, the Netherlands,
{w.r.j.van.breda, m.hoogendoorn}@vu.nl, j.pastor@student.vu.nl

[2] VU University Amsterdam, Department of Clinical Psychology, De Boelelaan 1081,
1081 HV Amsterdam, the Netherlands,
jruwaard@me.com, {j.a.g.j.asselbergs, h.riper}@vu.nl

Abstract. Mental health related problems are responsible for great sorrow for patients and social surrounding involved. The costs for society are estimated to be 2.5 trillion dollar worldwide. More detailed data about the mental states and behaviour is becoming available due to technological developments, e.g. using Ecological Momentary Assessments. Unfortunately this wealth of data is not utilized: data-driven predictive models for short-term developments could contribute to more personalized interventions, but are rarely seen. In this paper we study how modern machine learning techniques can contribute to better models for predicting short-term mood in the context of depression. The models are based on data obtained from an experiment among 27 participants. During the study frequent mood assessments were performed and usage and sensor data of the mobile phone was recorded. Results show that much can be improved before fine-grained mood prediction is useful within E-health applications. Subsequently important next steps are identified.

1 Introduction

Mental health problems have a high impact on the lives of patients, their social surrounding, and the society in general. It obstructs patients to learn, work or participate in society. Many affected therefore turn to professional help. The costs for society have been estimated at a startling amount of 2.5 trillion dollars per year worldwide in 2015, and will rise to 6 trillion dollars per year by 2030 [4]. These problems are the driving force behind health-related research that aims to provide more effective therapies for patients and their social environment. Major developments in mobile technologies offer new possibilities for mental health interventions, such as performing Ecological Momentary Assessments (EMA) more effectively. Using new technology you can more frequently assess the mental state of the user as well as the context in which it is measured. The context can be collected unobtrusively using sensors. Such measurements provide highly detailed insights into the behavior and mental state of the patient and can be a driver for more personalized and effective therapies.

Although studies that involve EMA are increasing in number, only very few studies try to fully take advantage of the wealth of data that results. Predictive modeling on a more detailed level, e.g. predicting mood level changes in terms of hours, are rarely seen while they can be a great driver for more real-time (semi) automated forms of therapy. Most predictive modeling endeavors focus on more long term predictions such as therapeutic effectiveness or long term recovery. There are some exceptions, such as [8] or [1]. However, these studies do not take advantage of more recent developments in the area of machine learning that can lead to more accurate results and as a consequence show relatively poor performance.

In this paper, we try to utilize sophisticated machine learning techniques to accurately predict the mood within the context of depression. Depression has a life-time prevalence of 17.1%, making it a major health problem [6], associated with morbidity, mortality, disability and psychological agony for the sufferers and their social surrounding [10]. Our starting point is a dataset collected among 33 participants in which sensory information from the smart phone of the participants has been collected as well as regular self-ratings of the mood. The current focus for the modeling is on predicting the current mood state of individual participants based on their individual histories as well as their current measurements using the smart phone. From the raw dataset, we derive a set of attributes following [8]. We then explore several ways to predict the aforementioned mood, specifically by using time-series techniques, dynamic time warping techniques, and a number of state-of-the-art machine learning techniques to see whether better techniques can contribute to more accurate predictions.

This paper is organized as follows. First, specific information about the data we will be using is discussed in Section 2, followed by the techniques we apply in Section 3. The results of our experiments are described in Section 4, and a discussion about the results can be found in Section 5.

2 Data

The data used in this research paper originates from the VU Unobtrusive Ecological Momentary Assessment pilot study data (for more info see [1]). First, the setup of the pilot study will be discussed, followed by the precise data that has been collected.

2.1 Pilot Study Setup

In the pilot study, measurements were performed over a period of approximately six weeks. A total of 33 participants were selected for the pilot of which 27 contributed enough data for meaningful analysis. The data consists of 76 variables and 1249 observations, running for 52 days. The data was obtained through two applications installed on the participant's smartphone: the eMate EMA application and the iYouVU application. The eMate EMA application prompts the user to rate their mood five times per day (at 09:00, 12:00, 15:00, 18:00 and 21:00) on a unidimensional scale as ”mood” on a scale from 1-10, and the iYouVU application is a sensor logger, which is active in the background, unnoticeable to the participant. The duration for which the participants logged data varies and manual input was not always provided. In case the participants expressed their mood multiple times a day, an average was calculated for each day.

2.2 Data Description

Table 1: The attributes that are present in the dataset in the form of averages per day. Most of the where T is the target feature, OF is an obtrusive feature, and UF is an unobtrusive feature

Attribute name	**Type**	**Information**	**Range**
mood	T	daily mean unidimensional to be predicted mood	[0,1]
mood.l1	OF	current mood, daily average	[0,1]
call.c1c - call.c5c	UF	number of calls made to top call contact 1 - 5	[0,1]
call.c1d - call.c5d	UF	duration of calls made to top call contact 1 -5	[0,1]
sms.c1c - sms.c5c	UF	number of SMS sent to top SMS contact 1 - 5	[0,1]
app.a1c - app.a5c	UF	number of times top app 1 - 5 was launched	[0,1]
app.a1d - app.a5d	UF	duration of use of top app 1 - 5	[0,1]
appCat.n	UF	app use frequency for each app category	[0,1]
appCat.sum	UF	app use duration for each app category	[0,1]
screen.duration	UF	mean screen-on moment (standardized)	[0,1]
screen.n	UF	screen-on frequency (standardized)	[0,1]
image.n	UF	number of photos taken on smartphone	[0,1]
accelerometer.high	UF	mean percentage of accelerometer.high data	[0,1]

For each participant, for each day an averaged mood value is present, which is referred to as an obtrusive feature attribute, because the participant manually needs to input their mood level over the day. The variables obtained through the iYouVU application are referred to as unobtrusive, because they are measured automatically. The values are aggregated per day. The number of calls and SMS messages to the top 5 contacts are measured as normalized frequency values. The duration of calls made to the top 5 contacts was measured, and the frequency and duration of the applications are measured as well. The appCat.n and appCat.duration are attributes representing the number of uses and duration for application categories. These variables are normalized per day between all categories of their respective class. The categories include: android, books, browser, business, education, entertainment, game, life style, email, music, news, productivity, social, tools, transportation, and unknown. The following attributes are also included: the number of images taken (image.n), the average screen duration per screen-on moment (screen.duration) and the screen-on frequency (screen.n), which were all normalized within the participant. Another variable obtained through the iYouVU application is the average percentage of accelerometer data that is classified as “high” (accelerometer.high). A summary of the dataset attributes are shown in Table 1.

3 Method

In this section we describe the methodology we have followed for generating our predictive models. We begin by discussing the steps taken to prepare the data: the process of derivation of additional attributes on top of the attributes described in Section 2 and the imputation of missing values. Then, we describe the techniques deployed on the dataset.

3.1 Data Preprocessing

Attribute Engineering The basic attributes which have directly been based on measurements in the pilot have already been discussed in Section 2. While these attributes are certainly useful, new information that can be derived from these attributes could potentially also be beneficial. We therefore constructed a number of additional attributes, that we included in a separate analysis. One attribute that was added concerns the weekday, which was calculated by using the already available data/time attribute. Another two attributes were the sum of call request and/or SMS to the top 5 persons for each day, and the sum of application usage of applications belonging to the top 5 used applications each day.

Missing value imputation The dataset contains several missing values for both the measurements of the smart-phone sensors as well as the EMA data. We consider data as missing if there is no single measurement on a day (the granularity of measurements considered in this case). There are a number of ways to deal with missing data; time points with missing values can either be removed or the missing values can be imputed. As already mentioned in the data description, the number of measurement points is quite limited. Therefore, the decision was made to only remove the time points of which the variable mood was missing, and impute the other values.

There are a total of 1249 time points, where for 1224 the mood variable is available. 1099 observations are without any missing values. We chose to impute the missing values with the mean of the variable per participant, using all observations of that individual.

3.2 Predictive Models

Different methods can be used to make numerical predictions. The decision for which predictive models to use in this experiment was made based on their properties. Predictive models we considered preferably are state-of-the-art algorithms, are known to perform well, can be used for regression, do not require much parameter tuning and do not take an excessive amount of time to train models. On a higher level the rationale behind selecting these predictive models for comparison is that we want to apply techniques that (1) only consider trends in the target based on its previous value (i.e. predict future mood based on past developments of mood): *mood only*, (2) try to use previously seen participants and predict based on their data: *similarity*, and (3) that use the feature attributes and prior values of mood for predicting future mood: *full predictive modeling*. Each of the three options mentioned above are explained in more detail below.[3]

Mood only For predicting mood exclusively using the mood series for each individual, we start with time series modeling using Auto Regressive Integrated Moving Average (ARIMA). The ARIMA model explores if the mood signal is a 'stationary' signal, i.e. the model tries to find statistical properties that make the mood series constant over time.[4]

Similarity To explore similarities in the mood series the Dynamic Time Warping (DTW) algorithm was applied. This algorithm compares two time-series for similarities, even if they differ in terms of time or speed. Such similarities could possibly exist between different individuals related to their mood series, and then possibly be used for predicting mood of individuals with highly similar mood series patterns. Since the DTW algorithm cannot deal with missing values, these values were imputed using the average of the variable before and after the missing value. This imputed sequence was normalized after which the algorithm was applied.[5]

Full predictive modeling For full predictive modeling we use two techniques: Support Vector Machines (SVM) and Random Forests (RF). SVM generates hyperplanes in a multidimensional space for purpose of classification or regression. Our implementation of SVM uses a gaussian radial basis function for the regression task.[6] The RF is a technique that averages the performance of a number of decision trees. The individual decision trees are trained on different parts of the dataset, and together do not suffer from overfitting problems.[7]

3.3 Setup and measurements

Mood only For each participant the ARIMA models use the whole series as training data. This first step, i.e. to reproduce the data, is important because we want to know the potential power of the method. Thus, the time series model has maximum potential to find statistical attributes to explain the dynamics for each individual. After this, for each individual the tuned model is fed the data related to each time point in an attempt to predict the target variable (the mood value) for each next time point. The final MSE for this method is calculated by taking into account the MSE for all individuals and all time points.

Similarity The similarity measure uses the whole series as training data as well, because we want to see the extent of similarities that are present between participants in the most ideal circumstance. In such a way we can better assess the potential of similarity measures for predicting mood series. Regarding the

[3] All models were constructed making use of R [11].

[4] For fitting the ARIMA model the Forecast package [3] was used (AICc with a correction for finite sample sizes)

[5] For using this technique the DWT package [2] was used, supported by [12] to allow for open begin/end comparisons. The clustering of time series was done making use of the TSClust package [9].

[6] For implementation we used the kernlab package [13].

[7] For implementation we used the randomForest package [7].

evaluation setup however, as is discussed in Section 4, we do not believe this has much potential yet. Therefore no evaluation setup is necessary for this method. Related to the settings of the algorithm, the ”Sakoe-Chiba” band was used with a maximum window size of 6. The hierarchical clustering that was applied used the complete-linkage clustering method.

Full predictive modeling For each participant the SVM and RF methods are trained based on the past days that are available and predict the target variable, the mood value, for the next time point. This way, as time increases, more training cycles become available for the machine learning method to train on. So, starting on day three, each model has two training cycles, predicting the mood for the current day; on day four, each model has three training cycles, and so on.The final MSE for each method is calculated by taking the MSE for all individuals for all time points. Due to this setup, the methods tend to have better performance as more training data cycles become available over time. The configuration for learning the algorithms is shown in Figure 1.

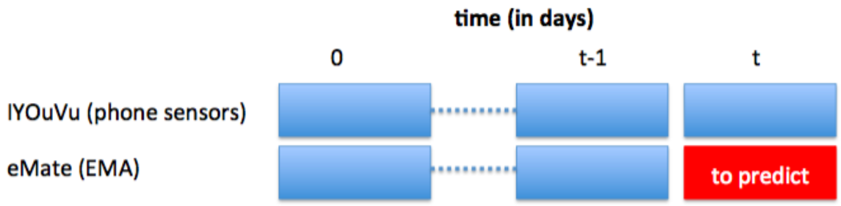

Fig. 1: Overview learning setup

Concerning the settings of the methods; for the SVM method, the epsilon regression was applied using a radial kernel, with the cost of constraint violation set to .5, epsilon to .1 and the sigma was determined by the built-in hyperparameter estimation heuristic function signet [5]; and for the RF method, the number of variables randomly sampled as candidates at each split was set to the number of columns divided by 3, the number of trees to grow to 500. Because fitting 1170 models (without parameter tuning) takes a considerable amount of time, no further parameter tuning on a per model basis was done as of yet, as this would further increase the computation times.

4 Results

As discussed in Section 3, we applied three types of methods: one where only the mood is considered for predicting mood for the next time point, one where we explore similarity analysis methods for predicting mood, and one where we apply full predictive modeling methods, namely RF and SVM. To compare the performance of the methods we added a benchmark method which simply predicts the same value as the average mood of all available past days.

4.1 Mood only

We start with time series analysis by deconstructing the univariate mood series of participants by representing them as ARIMA series. In this case we started with exploring how well the ARIMA is able to reproduce the trends seen for the mood as this is critical as a first step towards prediction. For exemplary purposes in Figure 2 the fit of the ARIMA models on the mood series of participant 1 and 5 is shown.

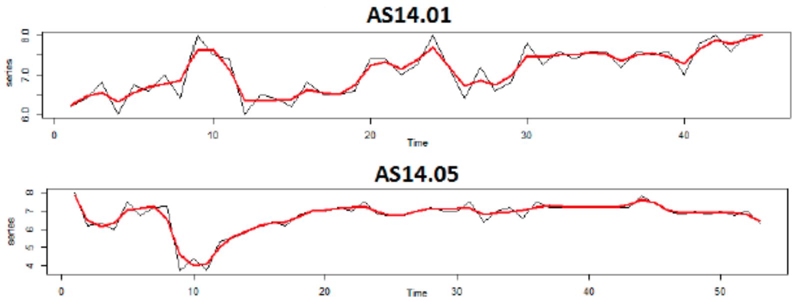

Fig. 2: Participant 1 and 5 - Best fit of the ARIMA models on the participant mood series.

Based on being trained on forehand on the full mood series for each individual. the ARIMA model produces an MSE of 0.475 for the reproduction of the mood patterns. Since this is even below the naive benchmark explained before (0.442) we conclude that using mood only is not enough to predict mood series accurately for individuals.

4.2 Similarity

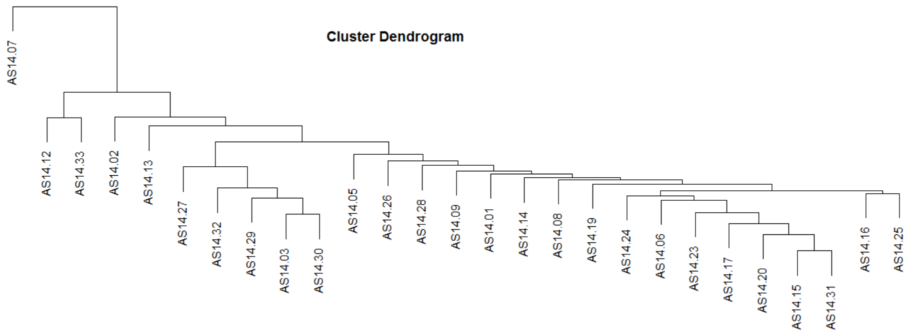

Fig. 3: Dendogram of Clustered time series

Next, we applied Dynamic Time Warping to see whether patients exhibit similar patterns and could potentially be used to predict the mood of unseen patients. The application of this approach resulted in a matrix with associated similarities between the different participants. We applied hierarchical clustering on this matrix [8]. The resulting dendogram can be found in Figure 3. It may be concluded that the mood series of participant 7 is quite unlike that of the other participants. Furthermore the dendogram provides insight into how the participants are clustered and provides a quick view into how a participant compares to others participants. An example of two similar time series and the resulting DWT path is illustrated in Figure 4, where the series of participant 16 and 24 are plotted. Although being an interesting analysis, as can be seen in Figure 4, based on the great variation in patterns and limited number of participants we did not consider this a viable option for prediction either.

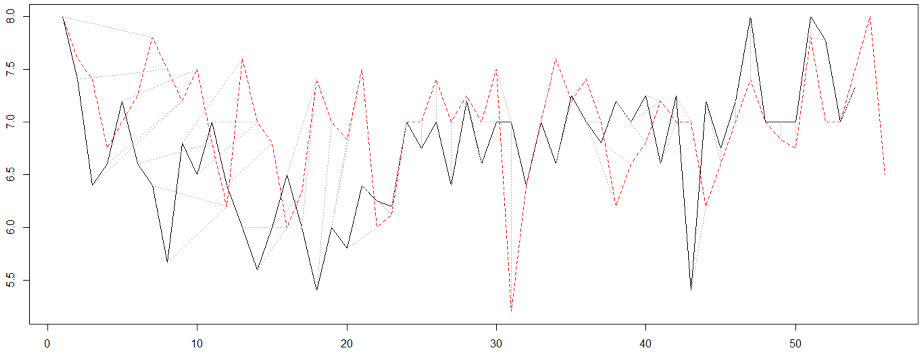

Fig. 4: Similarity DTW participants 16 and 24

4.3 Full predictive modeling

Finally, we applied the more traditional machine learning approaches. Here, we varied whether we used the added features (cf. Section 2 or not. Without these features, the SVM model obtains an MSE of 0.411, and the RF models scores 0.425. With the added features, the SVM models perform with a MSE of 0.410, and the RF models with a MSE of 0.420. In any case both methods are able to outperform the naive benchmark method, with an MSE of 0.442. Both SVM and RF did have a slight advantage with the added features. An overview of the performance of each of the methods is displayed in Table 2. Figure 5 shows how the MSE changes as more historical data is provided to the algorithm. It can be seen that predictive performance stabilizes after around 20 days of historical data. Note that the MSE is only measured for the day following the number of days considered as history.

[8] Specifically referred as UPGMA

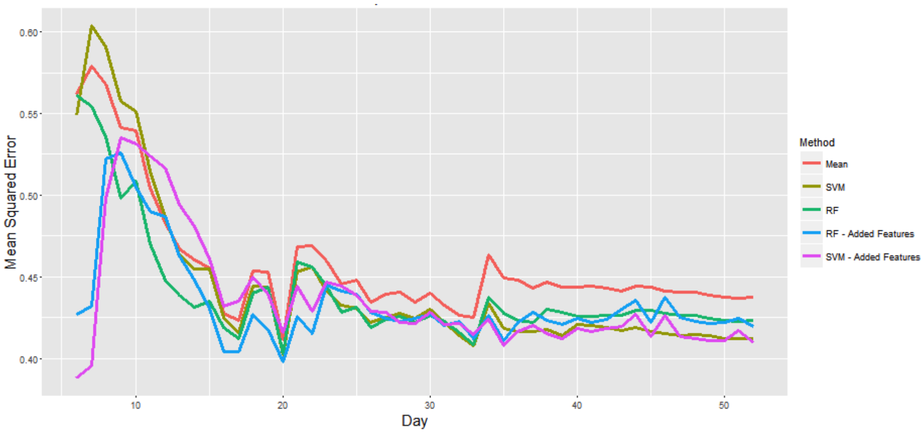

Fig. 5: The cumulative performance (MSE) over time for all participants. The x-axis shows the number of days of history considered.

5 Discussion

There is limited work on using sensor data to predict short-term developments of mood and related aspects, especially studies that have used machine learning approaches for this purpose. In this paper, we have therefore taken a more exploratory approach and applied a number of well-known machine learning techniques to a dataset collected in a small pilot study. We looked at the possibility to use time-series models to predict mood, where only the mood series itself was used without the measurements using the sensors on the phone. Thereafter we did include the unobtrusive measurements combined with dynamic time warping and more classical machine learning approaches.

Table 2: The performance of the different algorithms

Method name	**MSE**
Benchmark method (Naive)	0.441
SVM without added features	0.411
RF without added features	0.425
SVM with added features	0.410
RF with added features	0.420

Predicting mood based on past mood alone turned out to have low predictive performance. Mood generally does not seem to have intrinsic statistical properties that can explain the dynamics of mood to a large degree. Also, looking at similarities between mood series of different individuals, it was found that certain pairs of individuals do share more similarities than others, but we found the similarities too limited to be of use for prediction. Finally, the predictive models using SVM and RF that leveraged all data did result in better performance compared to the naive benchmark and the time-series models. This finding indicates that contextual data about the individual is useful and needed to increase predictive performance, but, given the attributes we had to our disposal, is not yet at the level to meaningfully employ within an E-health application. We think that the focus should be on finding the most relevant attributes that highly correlate with the dynamics of the target variable (in our case short-term mood), and think about how to optimally measure such attributes. Only then a sharp increase in predictive performance can be expected.

Next to finding new meaningful variables related to the target variable, we think preprocessing the data is an very important part of the process. The feature attributes we added in the preprocessing stage enabled a small increase in performance but we feel there is still a lot to gain.

References

1. J. Asselbergs, J. Ruwaard, M. Ejdys, N. Schrader, M. Sijbrandij, and H. Riper. Smartphone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study. *J Med Internet Res*, forthcoming.
2. T. Giorgino. Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, 31(7):1–24, 2009.
3. R. J. Hyndman and Y. Khandakar. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22, 2008.
4. T. R. Insel, P. Y. Collins, and S. E. Hyman. Darkness invisible: The hidden global costs of mental illness. *Foreign Aff.*, 94:127, 2015.
5. A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.
6. R. C. Kessler, K. A. McGonagle, S. Zhao, C. B. Nelson, M. Hughes, S. Eshleman, H.-U. Wittchen, and K. S. Kendler. Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey. *Archives of general psychiatry*, 51(1):8–19, 1994.
7. A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
8. R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. MoodScope: Building a Mood Sensor from Smartphone Usage Patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 389–402, New York, NY, USA, 2013. ACM.
9. P. Montero and J. A. Vilar. TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2014.
10. C. J. Murray and A. D. Lopez. Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. *The Lancet*, 349(9064):1498–1504, 1997.
11. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.
12. P. Tormene, T. Giorgino, S. Quaglini, and M. Stefanelli. Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine*, 45(1):11–34, 2008.
13. A. Zeileis, K. Hornik, A. Smola, and A. Karatzoglou. kernlab-an s4 package for kernel methods in r. *Journal of statistical software*, 11(9):1–20, 2004.

View publication stats