Gated Hidden Markov Models for Early Prediction of Outcome of Internet-Based Cognitive Behavioral Therapy

Negar Safinianaini[1(✉)], Henrik Boström[1(✉)], and Viktor Kaldo[2,3(✉)]

[1] School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
{negars,bostromh}@kth.se
[2] Department of Psychology, Faculty of Health and Life Sciences, Linnaeus University, Växjö, Sweden
[3] Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, and Stockholm Health Care Services, Stockholm County Council, Stockholm, Sweden
viktor.kaldo@ki.se

Abstract. Depression is a major threat to public health and its mitigation is considered to be of utmost importance. Internet-based Cognitive Behavioral Therapy (ICBT) is one of the employed treatments for depression. However, for the approach to be effective, it is crucial that the outcome of the treatment is accurately predicted as early as possible, to allow for its adaptation to the individual patient. Hidden Markov models (HMMs) have been commonly applied to characterize systematic changes in multivariate time series within health care. However, they have limited capabilities in capturing long-range interactions between emitted symbols. For the task of analyzing ICBT data, one such long-range interaction concerns the dependence of state transition on fractional change of emitted symbols. Gated Hidden Markov Models (GHMMs) are proposed as a solution to this problem. They extend standard HMMs by modifying the Expectation Maximization algorithm; for each observation sequence, the new algorithm regulates the transition probability update based on the fractional change, as specified by domain knowledge. GHMMs are compared to standard HMMs and a recently proposed approach, Inertial Hidden Markov Models, on the task of early prediction of ICBT outcome for treating depression; the algorithms are evaluated on outcome prediction, up to 7 weeks before ICBT ends. GHMMs are shown to outperform both alternative models, with an improvement of AUC ranging from 12 to 23%. These promising results indicate that considering fractional change of the observation sequence when updating state transition probabilities may indeed have a positive effect on early prediction of ICBT outcome.

Keywords: Hidden Markov Models · Expectation Maximization · Depression · Internet-based Cognitive Behavioral Therapy

D. Riaño et al. (Eds.): AIME 2019, LNAI 11526, pp. 160–169, 2019.
https://doi.org/10.1007/978-3-030-21642-9_22

1 Introduction

Depression affects about a hundred million people worldwide and it is estimated to reach second place in the ranking of Disability Adjusted Life Years for all ages in 2020 [15]. It is vital to consider depression an issue of public health importance, thereby prompting effective treatment of the patients and minimizing the disease burden [15]. Internet-based Cognitive Behavioral Therapy (ICBT) is an effective treatment for depression [22]. Machine learning can be used to solve different computational challenges in the analysis of ICBT, such as predicting patient adherence to depression treatment [20] and outcome prediction for obsessive-compulsive disorder [11].

A goal in the analysis of treatment outcome in ICBT for depression is to perform early predictions; thus, the patients with unsuccessful treatment can early on be detected and receive better care by the therapists. However, the accuracy of the predictions are also affected by at what time they are made; there is hence a trade-off for the psychologist to decide when to perform the early prediction. For example, if waiting one extra week gives better accuracy in predicting the final outcome, it may be preferred over deciding on a treatment earlier, based on a less accurate prediction. At the same time, waiting too long means less time to step in and adjust the treatment to better suit the patient [16].

The main motivation of this work is to explore a suitable machine learning method which improves the performance of early predictions on ICBT outcome for patients suffering from depression. We focus on graphical models, as they are interpretable, often easy to customize and allow for probabilistic modeling [1,23]. In particular, they allow for incorporating prior knowledge and handling missing data without imputation, through marginalization [2]. The latter is of particular importance as there are several, often unknown, reasons for why data is missing, and imputation may often not be appropriate in healthcare applications [9,21].

ICBT involves changes in human behavior; these have stochastic properties resulting in health state transitions. In this particular study, we have categorical observations (self-rated scores established by questionnaires) and a latent variable (treatment outcome) over time. As the state transitions can be modeled as Markov chains, Hidden Markov Models (HMMs) is a natural choice. However, one limitation of HMMs is the lack of context [24], which becomes a challenge when a state transition is dependent on the fractional change (defined as the difference between two values in time divided by the first value) of the observation sequence. We propose Gated Hidden Markov Models (GHMMs) as a potential solution to the problem. GHMMs extend standard HMMs by modifying the Expectation Maximization (EM) algorithm; for each observation sequence, the new algorithm regulates the transition probability update based on the fractional change, as specified by domain knowledge.

In the next section, we provide some notation and background on HMMs. In Sect. 3, we introduce the GHMMs. In Sect. 4, we evaluate and compare this approach to standard HMMs and a recently proposed approach, Inertial Hidden Markov Models (IHMMs) [13], on the task of early prediction of the outcome

of ICBT for treatment of depression. In Sect. 5, we discuss related work, and finally, in Sect. 6, we summarize the main findings and point out directions for future research.

2 Preliminaries

An HMM [2] is a statistical Markov model in which one observes a sequence of emitted symbols (observation sequence), but does not know the sequence of states the model went through to generate the observation sequence. The Markov property implies that the next state only depends on the current state. We define an HMM with N time steps, an observation sequence denoted as $X = \{x_1, \ldots, x_N\}$ containing N emitted symbols, and hidden states defined as $Z = \{z_1, \ldots, z_N\}$. An HMM has a parameter set, θ, which contains: initial probabilities, $p(z_1)$; transition probabilities, $p(z_n|z_{n-1})$; and emission probabilities, $p(x_n|z_n)$ where $n \in [1, N]$. The learning of the parameters of an HMM can be done by maximizing likelihood, using EM, which comprises two steps: the *E-step*, calculating the expected values; and the *M-step*, maximizing likelihood based on the expected values. Baum-Welch [2], shown in Algorithm 1, is an instance of EM suitable for HMMs. The E-step is done by calculating the marginal posterior distribution of a latent variable z_n, denoted as $\gamma(z_n)$, and the joint posterior distribution of two successive latent variables, $\varepsilon(z_{n-1}, z_n)$. In the M-step, θ is updated using $\gamma(z_n)$ and $\varepsilon(z_{n-1}, z_n)$. Forward and backward probabilities, $\alpha(z_n)$ and $\beta(z_n)$ [2], are used in the calculations of $\gamma(z_n)$ and $\varepsilon(z_{n-1}, z_n)$ as below. For details we refer to [2].

$$\gamma(z_n) = \frac{\alpha(z_n)\beta(z_n)}{p(X)} \qquad \varepsilon(z_{n-1}, z_n) = \frac{\alpha(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\beta(z_n)}{p(X)} \tag{1}$$

Algorithm 1. Baum-Welch

```
1: procedure LEARN(trainingData):
2: Initialise θ
3:     repeat
4:         for each X ∈ trainingData do
5:             E-step: calculate ε, γ in Equation (1)
6:             M-step: update θ using ε, γ
7:     until convergence
8: return θ
```

3 Gated Hidden Markov Models

Although HMMs are quite powerful, as demonstrated by their wide variety of applications, they have limitations in capturing long-range interactions between emitted symbols in the observation sequence; e.g. *Palindrome Language* [24].

Rather than considering more powerful (and less explored) models, we will in this proposal instead consider modifying the learning algorithm, i.e. Baum-Welch, to incorporate information regarding such long-range interactions through regulating the transition probabilities. In particular, we will consider global properties of the observation sequences, and when certain conditions are met, the algorithm, in the E-step, will be forced to set certain transition probabilities to zero. The latter can be thought of as gates being closed; hence the name Gated Hidden Markov Models (GHMMs).

Our algorithm, as presented in Algorithm 2, modifies Algorithm 1 by adding lines 6 through 8. Moreover, three additional input arguments (*policy*, *threshold*, *label*) and one new local variable (*change*) are added:

- *policy*: the rule defining how to calculate the fractional change
- *change*: the fractional change within an observation sequence, X, as calculated by *policy*.
- *threshold*: the specified threshold to compare with *change* (as determined by domain knowledge)
- *label*: the hidden state of the GHMM, which the algorithm regulates.

Algorithm 2. Modified Baum-Welch

```
1:  procedure LEARN(trainingData, policy, threshold, label):
2:  Initialise θ
3:      repeat
4:          for each X ∈ trainingData do
5:              Calculate γ in Equation (1)
6:              Calculate change by X and policy
7:  Gate    if change <threshold then
8:              p(z_n = label|z_{n−1}) = 0
9:              Calculate ε in Equation (1) by applying p(z_n = label|z_{n−1}) = 0
10:             M-step: update θ using ε, γ
11:     until convergence
12: return θ
```

Conceptually, the if-clause (line 7, Algorithm 2) represents the *Gate* concept. When the transition probability is set to zero, it means that the *Gate* is closed. Whenever this occurs, the update of θ in the M-step of EM is affected. The semantics of Baum-Welch is retained because the regulation only concerns the value of a transition probability and does not change any formulas calculated in the E-step or M-step. The algorithm can be viewed as updating the transition probabilities not only based on EM, but also based on the domain knowledge. Notice that the algorithm targets cases where the state transition is dependent on the fractional change of the observation sequence. The parameters *threshold*, *policy* and *label* may be customized for other situations, with similar types of data.

The worst-case time complexity of the modified algorithm is the same as for the original Baum-Welch algorithm; the worst case scenario of calculating the fractional change requires parsing the whole length of sequence, which results in that the original complexity is multiplied with a constant.

4 Empirical Investigation

4.1 Experimental Setup

Dataset. The data, based on the depression rating scale MADRS-S (Montgomery Åsberg Depression Rating Scale) [4], contain self-score replies to treatment questionnaires filled in by 2076 patients with depression and which have been assessed as suitable to, and willing to try, ICBT. The project, in which the data has been collected, has been approved by the regional ethical board in Stockholm (ref. no. 2011/2091-31/3, 2016/21-32 and 2017/2320-32).

The data points consist of ordinal values, ranging from 0 to 6, reflecting the severity of the mental state, as assessed by the patients themselves. The highest score represents the worst mental situation a patient can experience. The data is for each patient collected over 13 weeks, where for each week, the patient is requested to answer the same set of nine standardized questions. The data for the first week, week 0, is based on screening, before introducing the patient to ICBT, which contains the same questions. Week 0 is used when there is missing data regarding week 1 (week 1 corresponds to the pre-measurement week in [16]). Only patients that answered the questionnaires for the final week are included in the dataset (required for supervised learning). Let $q_i w_j$ denote the answer (a score from 0 to 6) to question i at week j. The observation sequence of $q_i w_j$s for each patient is assumed to be a merge of time-based (e.g. the step from $q_9 w_0$ to $q_1 w_1$) and event-based steps (e.g. the step from $q_1 w_0$ to $q_2 w_0$). For increasing the sequence size, which improves the learning of HMMs, we consider each step as a generic step in an HMM (regardless of whether it is event-based or time-based). The final observation sequence then becomes: $q_1 w_0, q_2 w_0, .., q_9 w_0, q_1 w_1, q_2 w_1, .., q_9 w_1, ..., q_1 w_{12}, q_2 w_{12}, .., q_9 w_{12}$.

The labels representing treatment outcome are "success" and "failure". Below, we show the rule concerning the class "success" based on clinical expertise [4] using the data from week 1 and week 12; the "failure" class does not satisfy the rule. The left inequality in Eq. 2 concerns the fractional change—called symptom reduction—being compared to the threshold of 50%; the right inequality in Eq. 2 defines the cut-off for the healthy score at the end of the treatment:

$$\frac{\sum_{i=1}^{9} q_i w_1 - \sum_{i=1}^{9} q_i w_{12}}{\sum_{i=1}^{9} q_i w_1} >= 0.50 \quad \vee \quad \sum_{i=1}^{9} q_i w_{12} <= 10 \tag{2}$$

The average symptom reduction over time and the frequency of the missing scores are shown in Fig. 1.

Experimental Protocol. This section explains the technical configurations of our experiments. For each of the considered algorithms, the same underlying structure is considered, consisting of the observation sequence and two hidden states, each corresponding to one of the two possible labels ("success" and "failure"). Here, we assume to know which hidden state corresponds to "success". For learning of the HMM parameters, we make the last latent variable observable by assigning the label to it, for each sequence X; we incorporate these changes into Baum-Welch.

For handling missing observations, marginalization is used based on [14]. We set the initial emission probabilities inspired by the prior knowledge; a patient's score for the "success" class has higher probabilities for the lower scores while for the "failure" class has higher probabilities for the higher scores.

The input parameters of Algorithm 2 are set to meet the requirements of the specific application of depression treatment using ICBT. The parameter *change* is set to be fractional change as defined in Eq. (2); *threshold* is set to 0.50 as in Eq. (2), and finally, *label* is set to "success". This means that if a patient's fractional change is less than 0.50, the transition probability of the outcome becoming "success" is set to zero (it is known which state transition probability to set to zero since the hidden state corresponding to "success" is known). The *Gate* here disallows EM to independently decide over the probability of treatment success if patients have insufficient fractional change, symptom reductions, according to the psychological measures. By this we have a hypothesis of reducing false negatives—the patients incorrectly predicted to belong to the "success" class—, which is critical for detecting that treatment is not successful.

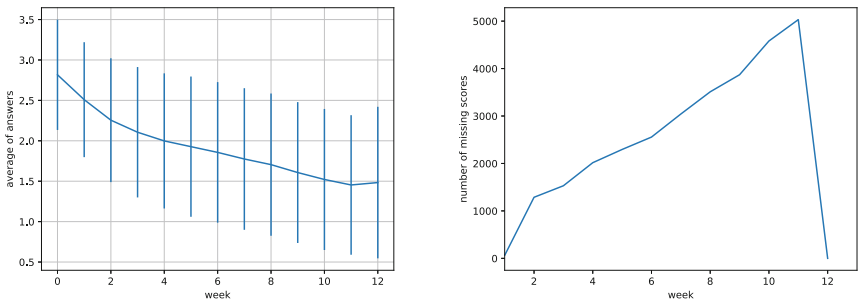

Fig. 1. On the left side, the average score for all patients through 13 weeks is shown, presenting the trend of symptom reduction. The vertical bars represent the standard deviation. To the right, the trend of missing scores is illustrated.

We compare the novel algorithm with HMMs and IHMMs. The latter regularizes the update of the transition matrix so that self-transitions, i.e., transitions to the same state as previous, have a higher probability than non-self-transitions. It is relevant to compare our algorithm with IHMMs since they satisfy the slow state transition property concerning a patient's behavior. We perform the comparison

on a separate test dataset and for each early prediction, data corresponding to the later weeks is withheld. The IHMM is trained with a set of values for the regularization parameter and the value resulting in the highest AUC in the validation set is chosen to be the regularization value. AUC, accuracy, precision and recall are used to evaluate the performance of the algorithms.

For the implementation of GHMMs, we refer to GHMMs.

4.2 Experimental Results

In Table 1, results are presented for GHMMs, HMMs and IHMMs regarding AUC, accuracy, precision and recall. The comparison is done for different early predictions with the earliest prediction taking place at week 5. This week, which corresponds to week 4 in [16], has shown to be the best week for measuring early change for ICBT [16]. GHMMs outperform HMMs and IHMMs with respect to AUC by between 12 to 23% and with respect to accuracy, with a probability threshold of 0.5, by between 2 to 8%. In Fig. 2, the performance comparison with respect to AUC is plotted. Evidently, GHMMs outperform the other models regarding all predictions which are up to 7 weeks before the final week.

Table 1. AUC, accuracy, precision and recall are compared for early predictions among three algorithms: HMMs; IHMMs, GHMMs.

% AUC

Week	HMM	IHMM	GHMM
12	67%	68%	91%
11	65%	65%	85%
10	65%	65%	85%
9	66%	67%	83%
8	63%	64%	80%
7	63%	63%	80%
6	66%	66%	78%
5	64%	64%	77%

% Accuracy (threshold 0.5)

Week	HMM	IHMM	GHMM
12	77%	77%	79%
11	69%	69%	77%
10	70%	70%	77%
9	71%	71%	77%
8	70%	70%	76%
7	72%	72%	74%
6	71%	71%	74%
5	70%	70%	73%

% Precision % Recall (threshold 0.5)

Week	HMM	IHMM	GHMM
12	99% 56%	99% 56%	98% 60%
11	90% 44%	90% 45%	90% 60%
10	88% 47%	88% 47%	88% 62%
9	84% 53%	84% 53%	86% 66%
8	78% 56%	79% 56%	81% 69%
7	78% 65%	78% 65%	77% 71%
6	75% 65%	75% 65%	76% 73%
5	73% 68%	73% 68%	73% 75%

Looking at precision and recall, in Table 1, it can be observed that GHMMs decrease false negatives more than the other algorithms for all weeks; confirming our hypothesis of reducing false negatives. Note that for all algorithms, when using probability threshold 0.50, precision gets higher but recall gets lower for later weeks; as shown in Fig. 2, however, week 12 dominates week 5, hence choosing a different threshold can lead to higher values for both precision and recall at later predictions.

5 Related Work

In medical applications, Markov models have been used to capture disease patterns regarding discrete mutually exclusive health states and the transitions between them over time. Markov models are useful in particular when the pattern involves clinical changes across the states; one clinical example being the

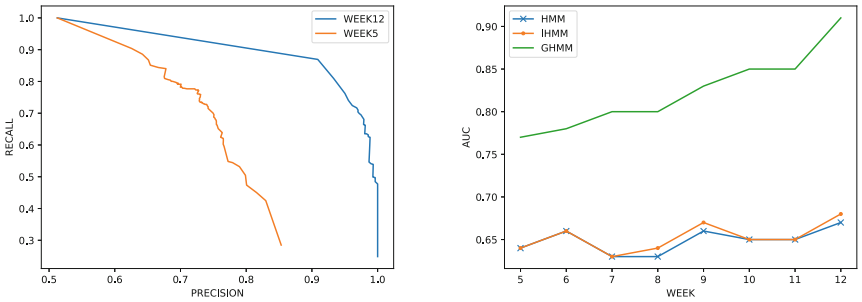

Fig. 2. On the left side, latest and earliest predictions by GHMMs, are compared concerning precision and recall. To the right, AUC of GHMMs, HMMs and IHMMs is compared for different early predictions.

progression of Alzheimer's Disease (AD) over time [6]. Shirley *et al.* [17] apply HMMs in alcoholism treatment analysis, by which different drinking behaviors are recognized. Assessment of preterm babies' health is another application of HMMs where the measurements are linked to state of health [10]. Capturing the quality of healthcare has been studied using HMMs for geriatric patient data by modelling quality as hidden states [12]. The clinical state of patients have also been estimated using infinite-HMM (an HMM with an unknown number of latent variables) [7].

Similar ideas to what have been proposed here, have also been used in integrating domain knowledge into machine learning; e.g. [3,8], where domain knowledge is applied in form of a framework or new components in the learning model. In contrast, GHMMs do not add any extra components to the model, as these may be expensive and complex. GHMMs instead apply domain knowledge through modifying the learning algorithm. Fung *et al.* [5] improve a binary classifier by incorporating two linear inequalities—so called knowledge sets—, corresponding to the classes, into the error minimization term of the classifier. We similarly use the linear inequalities between fractional change and the defined threshold as a constraint bundled in the optimization algorithm, EM.

Concerning context-sensitive HMMs for handling long-ranged interactions between symbols, in [24] an approach is proposed which stores symbols, emitted at certain states, in an auxiliary memory; the stored data serves as the context that affects the emission and the transition probabilities of the model. GHMMs also considers a symbol-based context, although without introducing extra components in HMMs. The early detection of neonatal sepsis has been studied using Autoregressive HMMs [18]; this work tackles HMMs' context limitation by introducing direct dependencies only between consecutive symbols. Similarly, GHMMs consider symbols dependencies but in a longer range.

6 Concluding Remarks

Standard HMMs have limited capabilities in capturing long-range interactions between emitted symbols in observation sequence. We introduce GHMMs as a remedy to this problem by which the learning of transition probabilities is regulated by the fractional change in observation sequence. This particular problem is motivated by the task of early prediction of ICBT outcome for depression. The approach is compared to standard HMMs and IHMMs, and GHMMs are shown to outperform both alternative models, with an improvement of AUC ranging from 12 to 23%, up to 7 weeks before ICBT ends. These promising results show that considering fractional change of observation sequence when updating state transition probabilities may have a positive effect on early prediction of ICBT outcome. These results, obtained through a collaboration project led by the Internet Psychiatry Clinic in Stockholm, indicate that GHMM may be a potentially effective tool in practice to improve predictions regarding ICBT [19].

The proposed approach can be applied and further tested in contexts of other psychological disorders and similar data types where the fractional change of an observation sequence should be allowed to affect state transitions. This work opens up for several different research paths, as there are still room for improvement, such as incorporating other forms of domain knowledge, considering additional data types, modelling missing values in the graphical model and combining the GHMMs with other machine learning and time series methods. Finally, regarding GHMMs, directions for future research include investigating soft thresholds and more complex gate mechanisms as well as techniques to avoid over-training of GHMMs.

References

1. Belgrave, D.: Machine learning for personalized health (ICML 2018). https://mlhealthtutorial.files.wordpress.com/2018/07/tutorial-ml-for-health1.pdf. Accessed 14 Aug 2018
2. Bishop, C.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York (2006)
3. Constantinou, A., Fenton, N., Neil, M.: Integrating expert knowledge with data in Bayesian networks: preserving data-driven expectations when the expert variables remain unobserved. Expert. Syst. Appl. **56**, 197–208 (2016)
4. Fantino, B., Moore, N.: The self-reported montgomery-asberg depression rating scale is a useful evaluative tool in major depressive disorder. BMC Psychiatry **9**(1), 26 (2009)
5. Fung, G., Mangasarian, O., Shavlik, J.: Knowledge-based support vector machine classifiers. In: NIPS (2003)
6. Green, C.: Modelling disease progression in Alzheimer's disease. Pharm. Econ. **25**(9), 735–750 (2007)
7. Hoiles, W., Van Der Schaar, M.: A non-parametric learning method for confidently estimating patient's clinical state and dynamics. In: NIPS (2016)
8. Kuusisto, F., Dutra, I., Elezaby, M., Mendonca, E., Shavlik, J., Burnside, E.: Leveraging expert knowledge to improve machine-learned decision support systems. AMIA Jt. Summits Transl. Sci. Proc., 87–91 (2015)

9. Kwak, Y., Yang, Y., Park, S.: Missing data analysis in drug-naïve Alzheimer's disease with behavioral and psychological symptoms. Yonsei Med. J. **54**(4), 825–831 (2013)
10. Lee, D., Roscoe, J., Russell, G.: Developing Hidden Markov Models for aiding the assessment of preterm babies- health. In: International Conference on Biomedical and Pharmaceutical Engineering, pp. 104–109 (2006)
11. Lenhard, F., et al.: Prediction of outcome in internet delivered cognitive behaviour therapy for paediatric obsessive compulsive disorder: a machine learning approach. Int. J. Methods Psychiatr. Res. **27**(1), e1576 (2018)
12. Mitchell, H., Marshall, A., Zenga, M.: Using the Hidden Markov Model to capture quality of care in lombardy geriatric wards. Oper. Res. Health Care **7**, 103–110 (2015)
13. Montanez, G., Amizadeh, S., Laptev, N.: Inertial Hidden Markov Models: modeling change in multivariate time series. In: AAAI Conference on Artificial Intelligence (2015)
14. Popov, A., Gultyaeva, T., Uvarov, V.: A comparison of some methods for training Hidden Markov Models on sequences with missing observations. In: 2016 11th International Forum on Strategic Technology (IFOST), pp. 431–435 (2016)
15. Reddy, M.: Depression: the disorder and the burden. Indian J. Psychol. Med. **32**(1), 1–2 (2010)
16. Schibbye, P., et al.: Using early change to predict outcome in cognitive behaviour therapy: exploring timeframe, calculation method, and differences of disorder-specific versus general measures. PLoS ONE **9**(6), e100614 (2014)
17. Shirley, K., Small, D., Lynch, K., Maisto, S., Oslin, D.: Hidden Markov Models for alcoholism treatment trial data. Ann. Appl. Stat. **4**(1), 366–395 (2010)
18. Stanculescu, I., Williams, C., Freer, Y.: Autoregressive Hidden Markov Models for the early detection of neonatal sepsis. IEEE J. Biomed. Health Inform. **18**(5), 1560–1570 (2014)
19. Titov, N., et al.: ICBT in routine care: a descriptive analysis of successful clinics in five countries. Internet Interv. **13**, 108–115 (2018)
20. Wallert, J., et al.: Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the u-care heart randomized controlled trial. J. Med. Internet Res. **20**(10), e10754 (2018)
21. Wartella, E.A.: The Prevention and Treatment of Missing Data in Clinical Trials. National Academies Press, Washington, D.C. (2010)
22. Williams, A., Andrews, G.: The effectiveness of internet cognitive behavioural therapy (ICBT) for depression in primary care: a quality assurance study. PLoS ONE **8**(2), E57447 (2013)
23. Xing, Z., Pei, J., Keogh, E.: A brief survey on sequence classification. ACM SIGKDD Explor. Newsl. **12**(1), 40–48 (2010)
24. Yoon, B., Vaidyanathan, P.: Context-sensitive Hidden Markov Models for modeling long-range dependencies in symbol sequences. IEEE Trans. Signal Process. **54**(11), 4169–4184 (2006)