Recent Context-aware LSTM for Clinical Event Time-series Prediction

Jeong Min Lee and Milos Hauskrecht

University of Pittsburgh, Pittsburgh PA 15260 USA
jlee@cs.pitt.edu, milos@pitt.edu

Abstract. In this work, we propose a novel clinical event time-series model based on the long short-term memory architecture (LSTM) that can predict future event occurrences for a large number of different clinical events. Our model relies on two sources of information to predict future events. One source is derived from the set of recently observed clinical events. The other one is based on the hidden state space defined by the LSTM that aims to abstract past, more distant, patient information that is predictive of future events. We evaluate our proposed model on electronic health record (EHRs) data derived from MIMIC-III dataset. We show that the combination of the two sources of information implemented in our method leads to improved prediction performance compared to the models based on individual sources.

Keywords: Recurrent Neural Network · Event time series prediction

1 Introduction

Successful modeling of complex multivariate event time series and their ability to predict future events is important for applications in various areas of science, engineering, and business. In clinical settings our ability to predict future events for a patient based on clinical events observed in past, such as past medication orders, past labs and their results, or past physiological signals can help us to anticipate the occurrence of a wide range of future events that would let health care practitioners intervene ahead of time or prepare resources to get ready for their occurrence. All of this can in turn improve the quality of patient care.

One of the challenges of modeling clinical event time series is their complexity, that is, clinical event time series for hospitalized patients may consist of thousands of different types of events corresponding to administration of many different medications, lab orders, arrivals of lab results, or various physiological observations, etc. This complexity may not fit very well standard Markov time series models [19] with either observed or hidden state and transition models.

To alleviate the event complexity problem we propose to develop a new more scalable event time series model based on the long-short-term-memory (LSTM) [14] that relies on two sources of information to predict future events. One source is derived from the set of recently observed clinical events. The other one is based on the hidden state space defined by the LSTM that aims to abstract past, more distant, patient information that is predictive of the future events. In the context of Markov state models, the next state in our models and the transition to the next state is defined by a combination of the recent state (most recent events) and the hidden state summarizing more distant past events.

In order to evaluate the proposed model, we use data derived from electronic health records (EHRs) of critical care patients in MIMIC-III dataset [16]. The clinical events considered in this work correspond to multiple types of events, such as medication administration events, lab test result events, physiological result events, and procedure events. These are combined together in a dynamically changing environment typical of intensive care units (ICUs) with patients suffering from severe life-threatening conditions.

Through extensive experiments on MIMIC-III data we show that our model outperforms multiple time series baselines in terms of the quality of event predictions. To provide further insights to its prediction performance we also divide the results with respect to different types of clinical events considered (medication, lab, procedure and physiological events), as well as, based on their repetition patterns, again showing the superior performance of our proposed model.

2 Related Work

2.1 Event-time series models

The majority of discrete time-series models are based on Markov processes [24, 25]. Markov process models rely on Markov property that assumes that the state captures all necessary information relating future and past. In other words, the next state depends only on the most recent state, and is independent of the past states. In this case the joint distribution of an observed sequence is modeled as chain of conditional probabilities: $p(y_1, y_2, ..y_T) = p(y_1) \prod_{t=2}^{T} p(y_t|y_{t-1})$

For Markov process models, the conditional probability defining a transition is parameterized by an $e \times e$ transition matrix where e denotes all possible states: $A_{i,j} = p(y_t = j|y_{t-1} = i)$. Standard Markov processes assume all states of the time series are directly observed. However, the states of many real-world processes are not directly observable. One way to resolve the problem is to define the state in terms of a limited number of past observations or features defined on past observations [31, 12, 11].

Hidden state models. Another is to use Hidden Markov models (HMM) [29] that introduce hidden states z_t of some dimension d. Now the observations y_t is defined in terms of the hidden states and an $e \times d$ emission table B with components: $B_{i,j} = p(y_t = j|z_t = i)$. Briefly, the transition table A is used to update the hidden states and the emission table is used to generate observations.

HMM has been shown to reach good performance in many applications such as stock price prediction [10], DNA sequence analysis [15], and time-series clustering [28]. However, classic HMM model comes with drawbacks when applied to real-world time series: the hidden state space is discrete, and the transition model is restricted to transitions in between the discrete states. Linear dynamical models (LDS) [17] remedy some of the limitations by defining real-valued hidden state-space with linear transitions among the current and next hidden state. One problem with HMM and LDS models is that the dimensionality of their hidden state space is not known a priori. Various methods for hidden state space regularization, such as work by Liu and Hauskrecht [21, 22] for LDS have been able to address this problem.

Continuous time models. We would like to note that in addition to discrete time series models, the researchers have explored also methods permitting continuous time models. Examples are various version of Gaussian process models for predicting multivariate time series in continuous time, including those used for representing irregularly sampled clinical time series [20, 23].

Neural-based models. Recent advances in neural architectures and their application to time-series offer end-to-end learning framework that is often more flexible than standard time-series models. In neural-based approaches, the discrete time series are typically modeled using recurrent neural network (RNN) which provides a more flexible framework for modeling time-series. Similarly to HMM and LDS, RNN uses hidden states to abstract and carry information from past history but with more flexible hidden state defined by real-valued vectors and transition rules. At each time step, hidden state is updated given the previous time step's hidden states and a new information from the current time step's input. Although its limitations on vanishing and exploding gradient problems [13], its variants such as long short-term memory (LSTM) [14] unit and gated recurrent units (GRU) [2] allow wide adoptions in event time-series modeling. They have been applied to prediction and modeling time series [9, 1], vision [8], speech [7], and language [30] problems.

2.2 Clinical event time-series modeling

Modeling and prediction of discrete event time series in the healthcare area have been influenced greatly by advances in various neural architectures and deep learning. [3] used Skipgram [26] to represent and predict next visit in outpatient data. But they evaluated their model on the prediction task at the level of hospital visit, which can be of a very coarse granularity for real-world clinical applications that encompass event-specific time information. [4] modeled clinical time series with RNN and attention mechanism. However, the model is only able to perform binary classification on a whole-sequence level. Our model is able to predict fine-grained future event at the level of each time step of a sequence. [6] also used neural network models to predict the sequence of clinical events. In their approach, the patient pool was limited to patients with kidney failure and organ transplant. On the other hand, our model is tested and shows superior performances over baselines across general clinical time series that were not limited to a specific patient cohort.

3 Methodology

In this section, we first introduce state-space Markov and LSTM-based event time series models and then present our model combining the two models.

State-space Markov event prediction. Given an observed events sequence $\mathbf{y} = y_1, y_2, ..., y_T$, we can model $\mathbf{y}$ by defining a Markov transition model relating the current event state y_t with the next event states y_{t+1}. In this case, we assume the event space is formed by a multivariate binary vector reflecting the occurrence of many different events (encoded as 1) over some time-window. One way to parameterize the transition between two consecutive event states is to use a transition matrix W with a bias vector b. As we want to predict multivariate binary vector, we can use sigmoid function $\sigma(x) = \frac{1}{1+e^{(-x)}}$ as the output activation function:

$$\hat{y}_{t+1} = \sigma(W \cdot y_t + b) \tag{1}$$

LSTM-based event prediction. LSTM models are being successfully used to model time series with the help of hidden state vector, allowing one to summarize in the hidden state information from more distant past. At a glance, at each time step of a sequence, LSTM gets current (event) input and updates its hidden states. The hidden state then generates signals for the next hidden state, as well as, predictions for the occurrence of events in the next time-step.

In detail, at each time step t, events in the input sequence represented as multi-hot vector m_t is processed to a real-valued vector x_t through linear embedding matrix W^{emb}: $x_t = W^{(emb)} \cdot m_t$. Then, given processed input x_t and previous hidden states h_{t-1}, LSTM updates hidden states h_t:

$$\begin{aligned} f_t &= \sigma(W^{(f)} \cdot [h_{t-1}, x_t] + b^{(f)}) & i_t &= \sigma(W^{(i)} \cdot [h_{t-1}, x_t] + b^{(i)}) \\ o_t &= \sigma(W^{(o)} \cdot [h_{t-1}, x_t] + b^{(o)}) & \tilde{C}_t &= \tanh(W^{c} \cdot [h_{t-1}, x_t] + b^{(c)}) \\ C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t & h_t &= o_t \otimes \tanh(C_t) \end{aligned}$$

f_t, i_t, and o_t are forget, input and output gates and $\otimes$ denotes element-wise multiplication. With these parameters ready, we can update hidden states:

$$h_t = \mathrm{LSTM}(x_t, h_{t-1})$$

Future event occurrence prediction is generated through a fully-connected layer W^q with output activation function sigmoid:

$$\hat{y}_{t+1} = \sigma(W^{(fc)} \cdot h_t + b^{(fc)}) \tag{2}$$

This parameterization links to the state space based event predictor. When y_t of Eq. 1 is replaced to hidden states h_t, it becomes Eq. 2.

Recent context-aware LSTM-based event predictor. When properly trained, hidden states in LSTM can be sufficient to represent and model future behaviors of event time-series by abstracting dependencies of past and future events. However, to be trained properly, LSTM (or any deep-learning based models) requires large amounts of training instances. In the clinical domain, obtaining large amounts of clinical cases (e.g., rarely ordered medication or lab tests) is hard in general. This constraint may deter us to train LSTM for predicting rare clinical cases. Meanwhile, for certain clinical event category such as medications, the future occurrence of an event may highly depend on recent previous or current occurrence of the event type and incorporating this information may help to resolve the data deficiency constraint.

Therefore, to address the problem, we propose and develop an adaptive mechanism that refers to both abstracted information of past sequence through hidden states of LSTM and concrete information about event occurrences in very recent context window. Different from the preliminary LSTM-based output generation in Eq. 2 that only depends on abstracted hidden states of LSTM, we directly refer to recent event occurrence information. The recent event at the current time step t is in multi-hot vector m_t and it is incorporated into the model through a linear transformation to model:

$$b^{(u)} = W^{(s)} \cdot m_t + b^{(s)}$$

$b^{(s)}$ can be seen as additional bias term that reflects recent event occurrence information and final prediction for event occurrence is made as follows:

$$\hat{y}_{t+1} = \sigma(W^{(fc)} \cdot h_t + b^{(fc)} + b^{(u)})$$

The proposed predictor also can be seen as combining the LSTM based predictor with state-space based Markov predictor. Especially, in context of Markov state models, the next state in our models and the transition to the next state is defined by a combination of the recent state (most recent events) and the hidden state summarizing more distant past events.

Loss function. To measure the performance of the event prediction, $\mathcal{L}$ is defined as binary cross entropy between label vector y_t and prediction vector $\hat{y}_t$ over all sequences in the training set and $\mathbf{1}$ denotes a vector filled with 1s:

$$\mathcal{L} = \sum_t -[y_t \cdot \log \hat{y}_t + (\mathbf{1} - y_t) \cdot \log(\mathbf{1} - \hat{y}_t)]$$

Parameter learning. The parameters of the model is learned by back propagation through time (BPTT) [32] with adaptive stochastic gradient descent based optimizer [18]. Hyper-parameters are tuned by F1-score performances on validation set with following ranges: embedding ($W^{(emb)}$) size in $\{128, 256, 512\}$; hidden states size in $\{512, 1024, 2048\}$ and learning rate $= 0.005$ batch size $= 512$. To prevent over-fitting, early stopping and dropout ($p = 0.5$) are applied.

4 Experimental Evaluation

4.1 Clinical data

We test the proposed model on MIMIC-III, a clinical database generated from real-world EHRs of intensive care unit patients [16]. We extract 21,897 patients whose records are generated from Meta Vision system that is one of the systems used to create records in the MIMIC-III database. We extract patient in age between 18 and 99 and whose length of stay in ICU is between 3 and 20 days. We randomly split patients into the train, test, and validation sets with the ratio of 7:2:1 and generate multivariate event time-series by segmenting sequences with both input-window and future window with size $W = 24$. At the end of each input-window, its future-window is generated.

We consider the following types of events in our models: medication administration events, lab results events, procedure events, and physiological result events. **Medication administration events** indicate records of specific kind of medication administered to the patient. **Lab results events** indicate lab test and its results represented as normal, abnormal-high, or abnormal-low. **Procedure events** indicate records of procedures patient received during hospitalization. For medication, lab, and procedure event categories, we select those events observed in more than 100 different patients. **Physiological result events** consist of 23 cardiovascular, routine vital signs, respiratory, and hemodynamics signals selected by a critical care expert. Similarly to the lab result events, numeric physiological results are discretized to normal, abnormal-high, and abnormal-low. Table 1 shows the basic data statistics.

Table 1. Clinical data statistics by event categories

Category	Medication	Procedure	Lab test	Physio signal
Cardinality	136	79	1197	102
Num. of occurrences	803K	257K	4266K	8378K
Proportion of positive label	5.9%	3.2%	3.6%	83.1%

4.2 Evaluation metrics

We evaluate the quality of time series predictions using area under precision-recall curve (AUPRC) and area under the receiver operating characteristic curve (AUROC). Although AUROC is commonly used to present result for binary classification problems, it can provide misleading information when applied to highly imbalanced dataset. On the other hand, AUPRC provides more accurate profile on performances of models under such circumstances [5, 27]. As shown in Table 1, our dataset is severely skewed to negative examples. Therefore, we use AUPRC as our primary evaluation measurement.

4.3 Baseline models

We compare our proposed model to the dense logistic regression models defined upon the following inputs (predictors) :
Current Markov state (Markov) as defined in Eq. 1.
Binary History (LR-binary): Unlike the current Markov state information, this model considers the occurrence of all past events (not just the most recent one) and encodes them into one multi-hot vector.
Count History (LR-count): This model, similarly to Binary history, summarizes all past events (not just the most recent ones), but instead of multi-hot vector representation it uses a vector of event counts.
Current LSTM state (LSTM): The model uses the hidden state of the LSTM to summarize information from distant past important for prediction.

4.4 Results

All our evaluations were performed on the test set, that was not touched during the training and validation steps. Prediction results in Figure 1 summarize the performance of our model and baselines on 24-hour prediction window. The results show that our model outperforms all baselines in terms of both AUROC and AUPRC statistics. Moreover, the Markov state model is better than pure LSTM in terms of AUPRC. This shows the information from the most recent time window is most of the time the most important source for predicting the next step events. This is not surprising given the fact that many events (such as drug administrations or lab orders) are repeated every 24-hours, hence once they are observed they are most likely to occur also in the next time window.

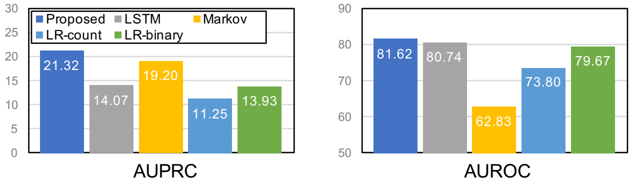

Fig. 1. Overall time-series prediction results on the 24-hours window segmentation

To verify the above reasoning, and to provide further insights into the predictive performance of our models, we break the above results by considering separately predictions when the same events occurred in the previous time step and when they did not. We refer to these as to repetitive and non-repetitive patterns. The results are given in Figure 2. From the results, we can clearly see that predicting non-repetitive events is significantly more difficult than predicting repetitive ones. However, despite this, we can also see that our model consistently outperforms other baselines across both repetitive and non-repetitive scenarios. Remarkably for non-repetitive event prediction, our model's AUROC is 32% higher than average of all baseline models in AUPRC and 11% in AUROC.

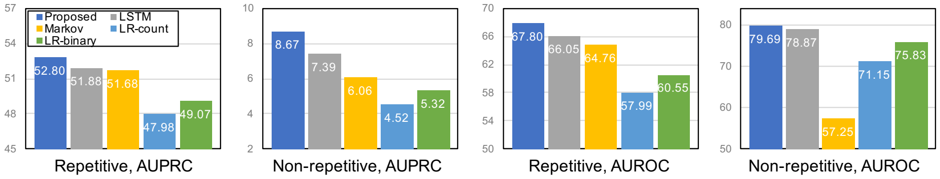

Fig. 2. Prediction results on repetitive and non-repetitive events

To analyze our results further, we next break the evaluation down by inspecting predictive performances of the models for the different event categories. The results are shown in Figure 3. Clearly, our model consistently outperforms baseline models across all event categories in both AUROC and AUPRC statistics.

So far, all our results were obtained by considering the window size of 24 hours. Next, we investigate the predictive performance of the models by varying the prediction window size. More specifically we will consider the window size W of length 6, 12 and 24 hours. Due to space limits, we will consider and compare the methods only using AUPRC statistics. As shown in Figure 4, our model shows superior performance across all time-resolutions.

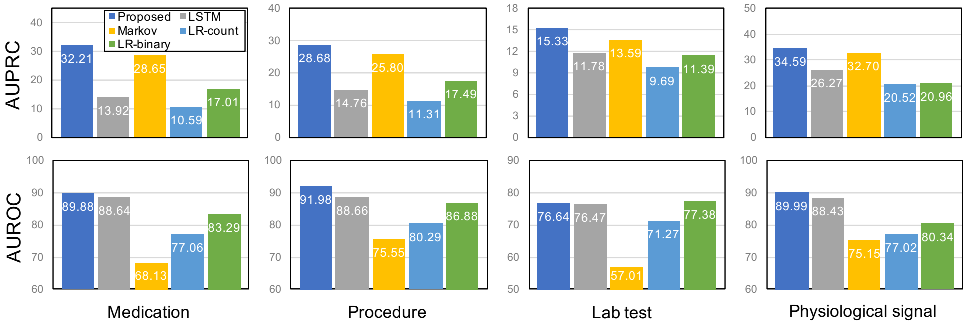

Fig. 3. Prediction results by the event type category

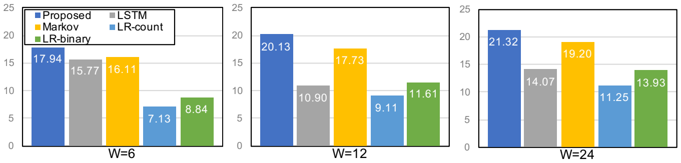

Fig. 4. AUPRC prediction statistics for the different window sizes

To dig deeper into the time segmentation results, in Figure 5 we show the predictive performance of lab test and physiological result events. We can see that on lab test event prediction, our model dominates at larger window sizes ($W = 12, 24$): it outperforms baseline models by 27%. In smaller window size ($W = 6$), the LSTM performs slightly better than ours by 2%. On physiological event prediction, our model surpasses all baselines across all time resolutions.

Interestingly, on lab event prediction, overall predictability is high at $W = 24$ and deteriorates for smaller window sizes. This reflects the recurrent characteristic of lab events at a cycle of 24 hours, that is, lab tests and their results are ordered and observed most of the time once daily. Inversely, the overall predictability of physiological events decreases with increasing window length. It indicates a recurrent characteristic of clinical events but in different recurring interval that is much shorter. Most physiological result events are automatically generated from bedside monitoring devices at short intervals, typically at a scale of seconds to minutes. Therefore, the variability of observation on a time series generated from smaller windows should be less than those of larger windows. Hence, overall predictability on smaller time resolution is consistently higher than larger ones as seen in Figure 5.

5 Conclusion

In this work, we show the importance of two sources of information for event-time series modeling. One source is derived from the set of recently observed clinical events and the other is based on the hidden states of LSTM that aims to abstract past, more distant, patient information that is predictive of future events. We show that the combination of the two sources of information implemented in our

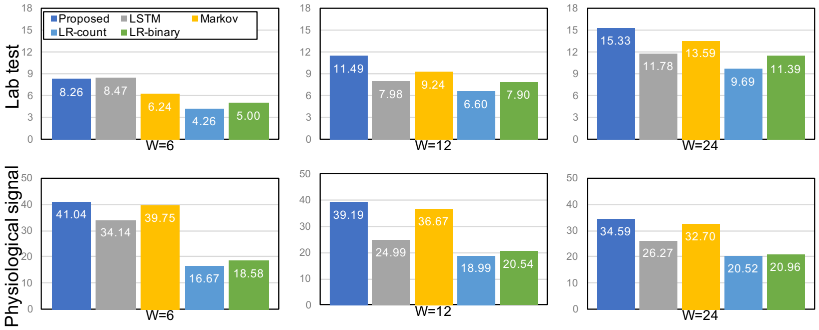

Fig. 5. Prediction result for lab and physiological events for the different window sizes

method leads to improved prediction performance on MIMIC-III clinical event data when compared to models that rely only on individual sources.

Acknowledgement. The work in this paper was supported by NIH grant R01GM088224. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. Chen, P.A., Chang, L.C., Chang, F.J.: Reinforced recurrent neural networks for multi-step-ahead flood forecasts. Journal of Hydrology **497**, 71–79 (2013)
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Choi, E., Bahadori, M.T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., Sun, J.: Multi-layer representation learning for medical concepts. In: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
4. Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., Stewart, W.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in Neural Information Processing Systems (2016)
5. Davis, J., Goadrich, M.: The relationship between precision-recall and the ROC curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240. ACM (2006)
6. Esteban, C., Schmidt, D., Krompa, D., Tresp, V.: Predicting sequences of clinical events by using a personalized temporal latent embedding model (2015)
7. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning. pp. 1764–1772 (2014)
8. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015)
9. Han, M., Xi, J., Xu, S., Yin, F.L.: Prediction of chaotic time series based on the recurrent predictor neural network. IEEE transactions on signal processing **52**(12), 3409–3416 (2004)
10. Hassan, M.R., Nath, B.: Stock market forecasting using hidden Markov model: a new approach. In: Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on. pp. 192–196. IEEE (2005)

11. Hauskrecht, M., Batal, I., Hong, C., Nguyen, Q., Cooper, G.F., Visweswaran, S., Clermont, G.: Outlier-based detection of unusual patient-management actions: An ICU study. Journal of Biomedical Informatics **64**, 211–221 (2016)
12. Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G.F., Clermont, G.: Outlier detection for patient monitoring and alerting. Journal of Biomedical Informatics **46**(1), 47–55 (2013)
13. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
15. Hughey, R., Krogh, A.: Hidden Markov models for sequence analysis: extension and analysis of the basic method. Bioinformatics **12**(2), 95–107 (1996)
16. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3** (2016)
17. Kalman, R.E.: Mathematical description of linear dynamical systems. Journal of the Society for Industrial and Applied Mathematics **1**(2), 152–192 (1963)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Lipton, Z.C., Berkowitz, J., Elkan, C.: A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019 (2015)
20. Liu, Z., Hauskrecht, M.: Clinical time series prediction: Toward a hierarchical dynamical system framework. Artificial Intelligence in Medicine **65**(1), 5–18 (2015)
21. Liu, Z., Hauskrecht, M.: A regularized linear dynamical system framework for multivariate time series analysis. In: The 29th AAAI Conference on Artificial Intelligence. pp. 1798–1804 (2015)
22. Liu, Z., Hauskrecht, M.: Learning linear dynamical systems from multivariate time series: A matrix factorization based framework. In: SIAM International Conference on Data Mining (2016)
23. Liu, Z., Wu, L., Hauskrecht, M.: Modeling clinical time series using gaussian process sequences. In: SIAM International Conference on Data Mining (2013)
24. MacDonald, I.L., Zucchini, W.: Hidden Markov and other models for discrete-valued time series, vol. 110. CRC Press (1997)
25. McKenzie, E.: Ch. 16. discrete variate time series. Handbook of Statistics **21** (2003)
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
27. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One **10**(3) (2015)
28. Smyth, P.: Clustering sequences with hidden Markov models. In: Advances in neural information processing systems. pp. 648–654 (1997)
29. Stratonovich, R.L.: Conditional Markov processes. Theory of Probability & Its Applications **5**(2), 156–178 (1960)
30. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–12 (2014)
31. Valko, M., Hauskrecht, M.: Feature importance analysis for patient management decisions. In: International Congress on Medical Informatics. pp. 861–865 (2010)
32. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proceedings of the IEEE **78**(10), 1550–1560 (1990)