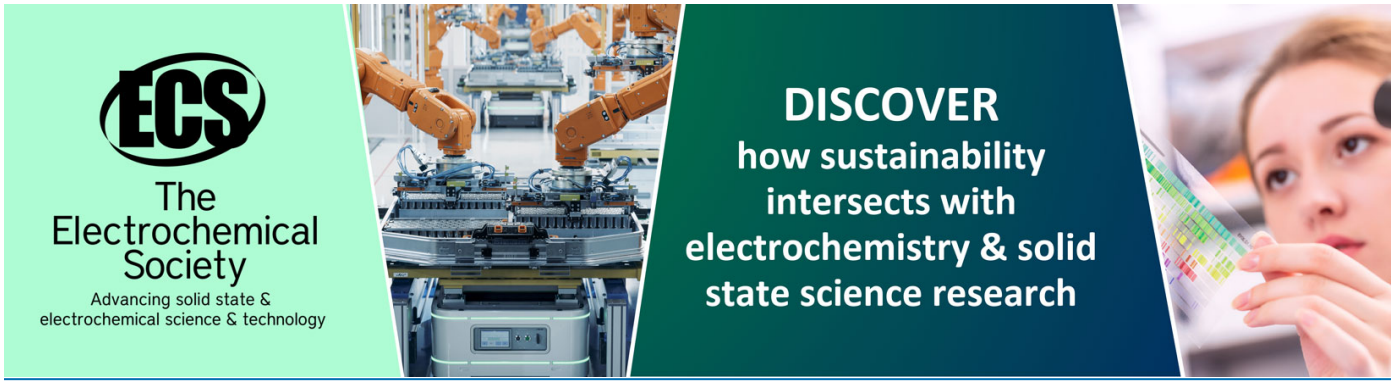
Imputation of Incomplete Motion Data Using Hidden Markov Models



IOP Conf. Series: Journal of Physics: Conf. Series **1210** (2019) 012151 doi:10.1088/1742-6596/1210/1/012151

Imputation of Incomplete Motion Data Using Hidden Markov Models

V E Uvarov, A A Popov and T A Gultyaeva
Novosibirsk State Technical University, Novosibirsk, Russia

Abstract. The modified Viterbi algorithm for imputation of incomplete sequences using hidden Markov models is presented. It is applied to the problem of imputation of incomplete motion data. It is shown that modified Viterbi algorithm outperforms imputation of gaps with the mean of neighbor observations.

1. Introduction

The problem of motion data analysis has many applications in the modern world. For example, smartphones, wearable and IoT devices may be provided with the function that determines if the device is used by someone who is not its owner and generates security warning. In other case, if wearable device is used by several family members, it can automatically determine the person who is using it at each moment. Other possible applications include military use (to make sure that equipment is used by the authorized user) or criminology use (to identify person who is using the device).

Unfortunately, sensor data is susceptible to distortion and information loss. For example, due to sensor temporary malfunction or some external noise. This is the reason why algorithms used for real-world applications of motion data analysis must be robust to such situations and be able to handle missing data appropriately.

One of the natural choices for the problems where data can be represented by sequences of observations are hidden Markov models (HMM) [1]. They were successfully applied to many practical problems, including signal processing and speech recognition [2]. However, the problem of using hidden Markov models in cases when the sequences are incomplete still remains poorly investigated. One of the attempts to apply HMMs to missing and unreliable data can be found in [3] where the authors used HMMs for noisy speech recognition problem. There was also some research on user identification by motion activity problem. For example, in [4] the authors were distinguishing individuals by their walking data, but HMMs weren't used. Attempt to use HMMs for the problem of user identification was demonstrated in [5] but the data was assumed to be complete and not noisy.

This paper is the continuation of research that is carried out at the department of theoretical and applied informatics of Novosibirsk state technical university [6-8].

This paper addresses the problem of imputation of motion activity data. The aim is to build a model from existing motion data of several individuals that can successfully impute new portions of incomplete data.

2. Hidden Markov model

2.1. Structure of hidden Markov model

Hidden Markov model (HMM) describes a random process which appears to be in one of the N hidden states at each time $t \in \{1, ..., T\}$ and transits to another or to the same state according to some transition probabilities. The states are hidden from the observer however they can be inferred from the observed sequences. In this paper, we consider HMM with continuous multivariate observation density, namely a mixture of multivariate normal distributions (often such HMM is called Gaussian HMM or GHMM).

We shall denote a hidden state of GHMM at time t as q_t, observation at time t as $\boldsymbol{o}_t$ and observation without specific time as $\boldsymbol{o}$. GHMM can be specified by initial state distribution

$$\Pi = \left\{ \pi_i = p\left(q_1 = s_i\right),\ i = \overline{1,N} \right\},$$

transition probabilities matrix

$$A = \left\{ a_{ij} = p\left(q_{t+1} = s_j \mid q_t = s_i\right),\ \ i, j = \overline{1,N} \right\}$$

and conditional multivariate distributions

$$B = \left\{ b_i\left(\boldsymbol{o}\right) = f\left(\boldsymbol{o} \mid q = s_i\right), i = \overline{1,N}, \boldsymbol{o} \in R^Z \right\}.$$

Here conditional multivariate distributions are mixtures of multivariate normal distributions

$$b_i(\boldsymbol{o}) = \sum_{m=1}^{M} \tau_{im} g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}),\ \ i = \overline{1,N}, \boldsymbol{o} \in R^Z,$$

where M – number of mixture components for each hidden state, $\tau_{im} \geq 0$ – weight of m-th mixture component in i-th hidden state ($\sum_{m=1}^{M} \tau_{im} = 1,\ i = \overline{1,N}$), μ_{im} – mean of normal distribution from m-th component of i-th hidden state, Σ_{im} – covariance matrix of normal distribution from m-th component of i-th hidden state and $g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}), \boldsymbol{o} \in R^Z$ – multivariate normal probability density function, i.e.

$$g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}) = \frac{1}{\sqrt{(2\pi)^Z \left|\Sigma_{im}\right|}} e^{-0.5(\boldsymbol{o}-\mu_{im})^T \Sigma_{im}^{-1} (\boldsymbol{o}-\mu_{im})}, \boldsymbol{o} \in R^Z.$$

Thus, some specific GHMM can be completely described by a set of parameters $\lambda = \left\{\Pi, A, B\right\}$ [1].

2.2. Decoding of complete sequences using Viterbi algorithm

For decoding of sequences using HMM, i.e. for inferring the most probable sequence of hidden states $Q = \left\{\hat{q}_1, \ldots, \hat{q}_T\right\}$ given the observation sequence $O = \left\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\right\}$ and HMM $\lambda = \left\{\Pi, A, B\right\}$ usually one would apply Viterbi algorithm [1] which is described below:

1. initialization:

$$\delta_1\left(i\right) = \pi_i b_i\left(\boldsymbol{o}_1\right),\quad i = \overline{1,N};$$

$$\psi_1\left(i\right) = 0;$$

2. induction:

$$\delta_t\left(j\right) = \max_{1 \leq i \leq N} \left[\delta_{t-1}\left(i\right) a_{ij} \right] b_j\left(\boldsymbol{o}_t\right),\quad j = \overline{1,N},\ \ t = \overline{2,T};$$

$$\psi_t\left(j\right) = \arg\max_{1 \leq i \leq N} \left[\delta_{t-1}\left(i\right) a_{ij} \right],\quad j = \overline{1,N},\ t = \overline{2,T};$$

3. termination:

$$\hat{q}_T = \arg\max_{1 \leq i \leq N} \left[\delta_T\left(i\right) \right];$$

4. path (state sequence) backtracking:

$$\hat{q}_t = \psi_{t+1}\left(\hat{q}_{t+1}\right),\quad t = \overline{T-1,\ 1}.$$

After execution of all the steps of the algorithm one would get the most probable sequence of hidden states: $Q = \left\{\hat{q}_1, \cdots, \hat{q}_T\right\}$.

3. Imputation of incomplete sequences using HMM

This section addresses the problem of incomplete sequence imputation i.e. the case when observation sequence $O = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ contain some missing observation or gaps which should be filled with the most probable values. We denote a missing observation as $\varnothing$.

3.1. Modified Viterbi algorithm

This proposed approach for imputation is based on modifying Viterbi algorithm so that it will be able to decode sequences with missing observations. The modified Viterbi algorithm is described below:

1. initialization:

$$\delta_1(i) = \begin{cases} \pi_i b_i(\boldsymbol{o}_1), & \boldsymbol{o}_1 \neq \varnothing \\ \pi_i, & \boldsymbol{o}_1 = \varnothing \end{cases}, \quad i = \overline{1, N};$$

$$\psi_1(i) = 0;$$

2. induction:

$$\delta_t(j) = \begin{cases} \max\limits_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right] b_j(\boldsymbol{o}_t), & \boldsymbol{o}_t \neq \varnothing \\ \max\limits_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right], & \boldsymbol{o}_t = \varnothing \end{cases}, \quad j = \overline{1, N}, \quad t = \overline{2, T};$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right], \quad j = \overline{1, N}, \; t = \overline{2, T};$$

3. termination:

$$\hat{q}_T = \arg\max_{1 \le i \le N} \left[\delta_T(i) \right];$$

4. path (state sequence) backtracking:

$$\hat{q}_t = \psi_{t+1}\left(\hat{q}_{t+1} \right), \quad t = \overline{T-1, \; 1}.$$

After execution of all the steps of the algorithm one would get the most probable sequence of hidden states: $Q = \left\{ \hat{q}_1, \cdots, \hat{q}_T \right\}$.

This most probable sequence of hidden states can be used for recovering the missing observations in the following way. The gap at a time t could with previously found hidden state $\hat{q}_t = s_{i^*}$ can be replaced by the most probable observation for that hidden state, i.e. $\hat{\boldsymbol{o}}_t = \arg\max\limits_{\boldsymbol{x} \in R^*} b_{i^*}(\boldsymbol{x}) = \arg\max\limits_{\boldsymbol{x} \in R^*} \sum\limits_{m=1}^{M} \tau_{i^* m} g(\boldsymbol{x}; \mu_{i^* m}, \Sigma_{i^* m})$. It should be obvious that when the mixture of normal distributions is used, the maximum value is reached when $\boldsymbol{x} = \mu_{i^* m^*}$ where $m^* = \arg\max\limits_m \left(\tau_{i^* m} \right)$, i.e. most probable value of $\boldsymbol{x}$ would be the mean of mixture component with the highest weight. Unfortunately, such strategy in some cases may lead to unwanted results. For example, HMM heavily overfits when trained on sequences imputed by this strategy according to our experiments. Another strategy is to replace gap with a value that is generated by a distribution that corresponds to i^*-th hidden state, i.e. $b_{i^*}(\boldsymbol{x})$ distribution. In our experiments for this paper we used the latter strategy.

3.2. Imputation with the mean of neighbouring observations

We compared the imputation using modified Viterbi algorithm with a standard method of imputation based on a mode of k neighboring observations.

After this imputation method is applied, some gaps may remain (e.g. such gaps that have k neighbors missing as well). That is why we apply this method of imputation again but the number of neighbors k is increased to match the length of the whole sequence T.

In this study we consider the 10 nearest neighbors of each gap (5 neighbors to the left and 5 to the left). This value was found empirically.

4. Evaluation

4.1. Motion Data Description

For algorithm evaluation we used "User Identification from Walking Activity" dataset available online. The dataset collects data from an Android smartphone positioned in the chest pocket. Accelerometer Data are collected from 22 participants walking in the wild over a predefined path. The dataset is intended for Activity Recognition research purposes. It provides challenges for identification and authentication of people using motion patterns [4].

The data for each of the participants is organized in tables with the following columns: time-step, x-acceleration, y-acceleration, z-acceleration. The accelerometer readings were acquired with frequency of 33Hz. Hence, the data for one user can be represented as a sequence of 3-dimensional vectors. The duration of measurements for one participant varies from 30 seconds to 11 minutes. A visualization of a short sample from data is presented in Fig. 1.

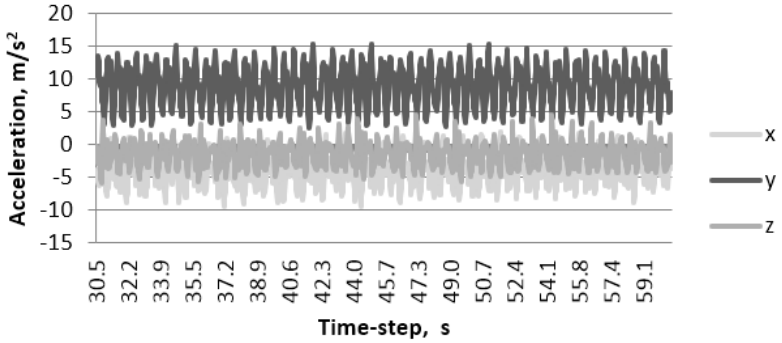


Figure 1. Short sample of accelerometer data

4.2. Computational Experiments

We trained one HMM for each of the participants. Each HMM had N=3 hidden states and M=3 mixture components of 3-dimensional (Z=3) Gaussian distributions. The number of states and mixture components were found empirically to provide the best accuracy with reasonable running time. Each sequence of 3-dimensional vectors was divided into subsequences of length T=100 (which roughly corresponds to 3 seconds of observations). We used 75% of randomly selected sequences from each class for training. Evaluation of models was performed on the remaining 25% of sequences. The final performance metric we used was the mean squared difference between actual and imputed observations.

The metric was calculated for various percent of missing observations in test sequences and for the two methods of imputation. The gaps places were chosen randomly for each sequence. The results are presented in Fig. 2.

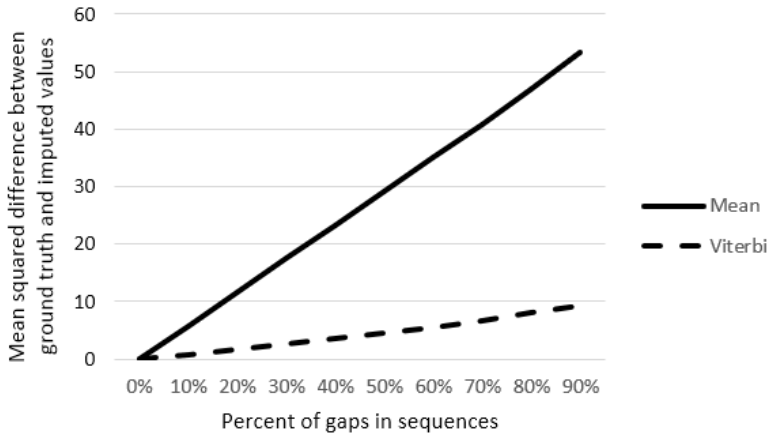


Fig. 2. Comparison of algorithms that perform imputation of incomplete sequences using HMMs trained on complete data

5. Conclusion

It can be seen from fig. 2 that imputation using modified Viterbi algorithm outperforms the standard imputation with the mean of neighboring observations. To sum up, based on the experiment results it can be recommended to use imputation using modified Viterbi algorithm to impute incomplete sequences when using HMMs for motion data recovery. It outperforms the standard method, easy to implement and brings no additional overhead to the imputation procedure.

References

[1] Rabiner L R 1989 A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition *Proceedings of the IEEE* vol 77 pp 257–285

[2] Gales M and Young S 2007 The Application of Hidden Markov Models in Speech Recognition *Signal Processing* vol **1** no 3 pp 195–304

[3] Cooke M, Green P, Josifovski L and Vizinh A 2001 Robust automatic speech recognition with missing and unreliable acoustic data *Speech Communication* vol **34** 3 pp 267–285

[4] Casale P, Pujol O and Radeva P 2012 Personalization and user verification in wearable systems using biometric walking patterns *Personal and Ubiquitous Computing* vol **16** no 5 pp 563–580

[5] Nickel C and Busch C 2013 Classifying accelerometer data via hidden Markov models to authenticate people by the way they walk *IEEE Aerospace and Electronic Systems Magazine* vol **28** no 10 pp 29–35

[6] Popov A Gultyaeva T and Uvarov V 2016 Training Hidden Markov Models on Incomplete Sequences *13th International Conference on Actual Problems of Electronic Instrument Engineering Proceedings (APEIE 2016)* vol **1** pp 317–320

[7] Uvarov V Popov A and Gultyaeva T 2017 Modeling multidimensional incomplete sequences using hidden Markov models *Proceedings of the International Workshop Applied Methods of Statistical Analysis Nonparametric approach (AMSA-2017)* pp 343–349

[8] Uvarov V E, Popov A A and Gultyaeva T A 2018 Recognition of incomplete sequences using Fisher scores and hidden Markov models *Journal of Physics: Conference Series, XI International scientific and technical conference Applied Mechanics and Dynamics Systems* vol **944** no 1