ITAT 2016 Proceedings, CEUR Workshop Proceedings Vol. 1649, pp. 172–178
http://ceur-ws.org/Vol-1649, Series ISSN 1613-0073, © 2016 T. Šabata, T. Borovička, M. Holeňa

Modeling and Clustering the Behavior of Animals Using Hidden Markov Models

Tomáš Šabata[1], Tomáš Borovička[1], and Martin Holeňa[2]

[1] Faculty of Information Technology,
Czech Technical University in Prague,
Prague, The Czech Republic
[2] Institute of Computer Science,
Czech Academy of Sciences,
Prague, The Czech Republic

Abstract: The objectives of this article are to model behavior of individual animals and to cluster the resulting models in order to group animals with similar behavior patterns. Hidden Markov models are considered suitable for clustering purposes. Their clustering is well studied, however, only if the observable variables can be assumed to be Gaussian mixtures, which is not valid in our case. Therefore, we use the Kullback-Leibler divergence to cluster hidden Markov models with observable variables that have an arbitrary distribution. Hierarchical and spectral clustering is applied. To evaluate the modeling approach, an experiment was performed and an accuracy of 83.86% was reached in predicting behavioral sequences of individual animals. Results of clustering were evaluated by means of statistical descriptors of the animals and by a domain expert, both methods confirm that the results of clustering are meaningful.

1 Introduction

A mathematical model that describes behavior is called behavioral model. In particular, we assume that patterns of animals' behavior are reflected in their behavioral models and the differences can be identified and analyzed on the individual level.

For each animal, a model that represents its behavior in a certain time period is created. A comparison of models of one animal from different periods of time can show changes in its behavior over time. Rapid changes in the behavior can indicate an important event or disorder.

Moreover, different animals can be compared based on their behavioral models. The behavioral model as a descriptor of the behavioral patterns can be used to group and classify the animals. Furthermore, characteristics of each group can be extracted and changes of behavioral patterns can be tracked.

2 Related work

The field of the behavioral analysis is very wide [2]. However, the paper focuses on a specific kind of behavioral analysis, behavioral analysis of animals as a sequence of states.

To create a sufficiently accurate behavioral model, an abstraction and simplification of the behavior is used. The most common abstraction of an animals' behavior is the abstraction as a sequence of states

The description of behavior as a sequence of states is a commonly used abstraction and simplification in modeling behavior [6, 13, 14]. This allows to create generalised and sufficiently accurate behavioral model. Each state of the sequence corresponds to an action. Actions are organized in a finite sequence [6, 13, 14, 17, 20]. The model is expected to be more accurate if actions are easily separable. Thus, these actions should be mutually disjoint and cover all activities that an animal can do. It means that the animal is doing exactly one action at a time.

An abstraction of an animal living in a closed environment is much simpler than for example an abstraction of the behavior of a human. Such animals can do only a limited number of actions. These actions are reactions to the internal state of the animal (hunger, thirst, ...), reactions to other animals' behavior or to the environment.

The most commonly used methods to model sequences are Hidden Markov Models [14], Dynamic Bayesian networks [16], Conditional Random Fields [12], Neural Networks [15], Linear regression model [4] or Structured Support Vector machines [22]. All these methods belong to methods of supervised learning. Even though, Hidden markov models can be also estimated in a semisupervised way [23]. There are also approaches that utilize unsupervised learning methods to deal with modeling of behavior. The Fuzzy Q-state Learning is an example of unsupervised method that was used to model humans' behavior. In this approach labels are created by the Iterative Bayesian Fuzzy Clustering algorithm [13].

3 Preliminaries

3.1 Hidden Markov Models

Influenced by [6, 14, 20], we have decided to use hidden Markov models for behavioral modeling. In this subsection, the principles of a hidden Markov model (HMM) will be recalled.

With each HMM, a random process indexed by time is connected, which is assumed to be in exactly one of a set of

N distinct states at any time. At regularly spaced discrete times, the system changes its state according to probabilities of transitions between states. Time steps associated with time changes are denoted $t = 1, 2, 3, \ldots$. The actual state at a time step t is denoted q_t.

The process itself is assumed to be a Markov chain, usually a first-order Markov chain, although a Markov chain of any order can be used. It is usually assumed that the Markov chain is homogeneous. This assumption allows the Markov chain to be described as a matrix of transition probabilities $A = \{a_{ij}\}$, which is formally defined in the Equation (1).

$$a_{ij} = P(q_t = y_j | q_{t-1} = y_i), \quad 1 \leq i, j \leq N \tag{1}$$

The simple observable Markov chain is too restrictive to describe the reality, however it can be extended. Denoting Y the variable recording the states of the Markov chain, a HMM is obtained through completing Y with a multivariate random variable X. In the context of that HMM, X is called 'observation variable' or 'output variable', whereas Y is called 'hidden variable'. The the hidden variable takes values in the set $\{y_1, y_2, ..., y_N\}$ and observable variable X takes values in the set $\{x_1, x_2, ..., x_M\}$.

We assume to have an observation sequence $O = o_1 o_2 ... o_T$ and a state sequence $Q = q_1 q_2 ... q_T$ which corresponds to the observation sequence. HMM can be characterized using three probability distributions:

1. A state transition probability distribution $A = \{a_{ij}\}$ which is formally defined by the Equation (1).

2. An probability distribution of observation variables, $B = \{b_{i,j}\}$, where $b_{i,j}$ is a probability of observation variables in state y_i and it is formally defined as (2).

$$b_{i,j} = P(o_t = x_j | q_t = y_i) \tag{2}$$

   The matrix B of the discrete variable can be described using $N \times M$ stochastic matrix, where M denotes number of possible values of X.

3. An initial state distribution $\pi = \{\pi_i\}$ is defined by

$$\pi_i = P(q_1 = y_i) \tag{3}$$

   When the initial state distribution is assumed to be stationary, then a initial state distribution is usually computed as the stationary distribution of the Markov chain described with matrix A by solving the Equation (4).

$$\begin{aligned} \pi A &= \pi \\ \sum_i \pi_i A_{ij} &= \pi_j \end{aligned} \tag{4}$$

With these three elements, the HMM is fully defined. The model is denoted $\lambda = (A, B, \pi)$. A HMM can be graphically depicted by Trellis diagram which is shown in Figure 1. The joint probability of O and Q, which corresponds to the trellis diagram, is described by the Equation (5). By means of the above notation, it can be formally described as in the Equation (6).

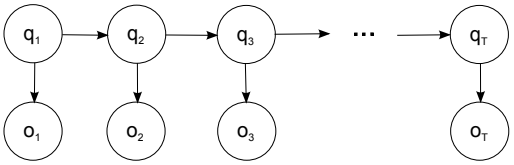

Figure 1: Trellis diagram of a HMM.

$$P(o_1, ..., o_T, q_1, ..., q_T) = P(q_1) P(o_1 | q_1) \prod_{k=2}^{T} P(q_k | q_{k-1}) P(o_k | q_k) \tag{5}$$

$$P(o_1, ..., o_T, q_1, ..., q_T) = \pi_{q_1} b_{q_1, o_1} \prod_{k=2}^{T} a_{q_{k-1}, q_k} b_{q_k, o_k} \tag{6}$$

3.2 Clustering approaches

Hidden Markov models are often used in cluster analysis of sequences [1, 9, 18]. Clustering sequences is more complex task than clustering feature vectors since infinite sequences are not in a Euclidean space and sufficiently long finite sequences are in a high-dimensional Euclidean spaces. Although there exist approaches how to measure distance between two sequences (Levenshtein distance, Hamming distance, longest common subsequence, etc.), however, they may not be always meaningful. It may be more reasonable to learn a HMM for each sequence and cluster the HMMs. A HMM is the random process that produces the sequence.

Due to the fact that HMMs' parameters lie on a nonlinear manifold, application of the k-means algorithm would not succeed [3]. In addition, parameters of a particular generative model are not unique, therefore, a permutation of states may correspond to the same model [3]. It means that different sequences may be generated by the same HMM.

There are already a few approaches to the cluster analysis by means of HMMs. One of them uses the spectral clustering with Bhattacharyya divergence [9]. Another commonly used dissimilarity measure in spectral clustering of HMMs is the Kullback-Leibler divergence [10, 21, 24].

The approach called variational hierarchical expectation maximization (VHEM) [3] is also based on spectral clustering. It directly uses the probability distributions of HMMs instead of constructing an initial embedding and besides clustering, it generates new HMMs. The newly generated HMMs are representatives of each cluster.

A disadvantage of all the mentioned methods is the assumption that observations are normally distributed and can be described using the Gaussian mixture model. A

distance measure between two HMMs with normally distributed observations can be computed in polynomial time. However, that assumption can be too restrictive for many real-life applications.

Dissimilarity measures of HMMs The definition of a distance between two HMMs is challenging since HMM consists of a Markov chain and a joint distribution of observations. In [5], the authors considered two distance measures applicable to HMMs λ and λ'. The first of them is a normalized Euclidean distance of the corresponding matrices B, B',

$$d_{ec}(\lambda,\lambda') = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}||b_{ik}-b'_{ik}||^2}, \tag{7}$$

where N denotes a number of states and M denotes a number of observations. The number of states of both models has to be identical $N = N'$. The second distance proposed in [5] is defined by

$$d_{mec}(\lambda,\lambda') = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\min_j\sum_{k=1}^{M}||b_{ik}-b'_{jk}||^2}. \tag{8}$$

A disadvantage of the considered distances is that they ignore the Markov chain. Therefore, there exist HMMs λ and λ' the distance between which is zero although the generated probability distributions P_λ and P'_λ are different. Consequently, both distance measures are not metrics, but only pseudometrics [11]. Although the distances ignore the Markov chain, it is agreed that the observation probability distribution matrix B is, in most cases, a more sensitive set of parameters related to the closeness of HMMs then the initial distribution vector π or the state transitions matrix A [10].

Another distance between HMMs λ, λ' is the Kullback-Leibler divergence (KL divergence)

$$d_{KL}(\lambda,\lambda') = \int_O \frac{1}{G(O)}\log\frac{P(O|\lambda)}{P(O|\lambda')}P(O|\lambda)dO, \tag{9}$$

where $G(O)$ is a function weighting the length of the sequence O. Two kinds of such weighting functions are commonly used. The function $G(O)$ is equal to the length of sequence if we measure the divergence between two HMMs that generate equally long sequences. It is equal to expected value of lenghts of sequences if we measure the divergence between two HMMs which generate various long sequences.

An analytic solution for integral in the Equation (9) exists for HMMs with Gaussian distributed observations [8], otherwise it can be calculated numerically. The time complexity of the numerical computation grows exponentially with the length of the sequence. Since we want to use it for long sequences we use an aproximation described in [5]. We assume to have an ordered set of output sequences $F = \{O_1,...,O_L\}$. Any sequence O could became l-th member of set with a probability proportional to $P(O|\lambda)$. Using Viterbi algorithm, the most likely sequence Q_{opt} of states can be computed by $P(Q_{\text{opt}},O|\lambda) = max_Q P(Q,O|\lambda)$. This leads to the KL divergence.

$$d_{\text{Vit}}(\lambda,\lambda') = \int_O \frac{1}{G(O)}\log\frac{P(Q_{\text{opt}},O|\lambda)}{P(Q'_{\text{opt}},O|\lambda')}P(O|\lambda)dO \tag{10}$$

We assume that:

- The most probable sequences of both hidden Markov models are equal for both HMMs. The assumption is reasonable if two HMMs are not too dissimiliar.
- The Markov chain is ergodic. A Markov chain is called ergodic if there is a nonzero probability to get from any state q to any other state q' (not necessarily in one move). A Markov chain with an ergodic subset (in particular, an ergodic Markov chain) generates sufficient long sequence.

With these assumptions, the KL divergence can be computed using following equation

$$d_{Vit}(\lambda,\lambda') = \frac{1}{G(O)}\log\frac{P(Q_l,O|\lambda)}{P(Q_l,O|\lambda')}P(O|\lambda)+\varepsilon, \tag{11}$$

where ε is an approximation error [5]. According to Subsection 3.1, the equation can be rewritten using the notation of HMMs as it is shown in the Equation (12).

$$\begin{aligned} d_{Vit}(\lambda,\lambda')-\varepsilon = &\frac{1}{G(O)}\sum_{t=1}^{T-1}\left(\log a_{y_t,y_{t+1}}-\log a'_{y_t,y_{t+1}}\right)+ \\ &\frac{1}{G(O)}\sum_{t=1}^{T}\left(\log b_{y_t,x_t}-\log b'_{y_t,x_t}\right) \end{aligned} \tag{12}$$

If a sequence O is long enough, the law of large number allows us to approximate Equation (12) as (13) [5].

$$\begin{aligned} d_{Vit}(\lambda,\lambda') \approx \widetilde{\delta}_{Vit} = &\sum_{i,j}a_{ij}\pi_i\left(\log a_{ij}-\log a'_{ij}\right)+ \\ &\sum_{i,k}b_{ik}\pi_i\left(\log b_{ij}-\log b'_{ij}\right) \end{aligned} \tag{13}$$

A disadvantage of the KL divergence is that it is not symmetric.

The Bhattacharyya divergence is also a commonly used metric for HMMs. For two probability distributions f and g, it is defined by the Equation (14). Similarly to the KL divergence, it can be easily computed if HMM has normally distributed observations [7]. Differently to the KL divergence, the Bhattacharyya divergence is symmetric.

$$d_B(f,g) = -\ln\left(\int\sqrt{f(x)g(x)}dx\right) \tag{14}$$

Using a chosen dissimilarity measure, a dissimilarity matrix can be created and used for clustering, in particular, in hierarchical and spectral clustering algorithms.

Spectral clustering uses similarities between objects represented as a weighted graph. There are three ways how to create such a graph:

1. **The ε-neighborhood graph.** In this way, all pairs of points with distances smaller than ε are connected. The graph is often considered as unweighted because its edges are based on the dichotomous property of distances at most ε.

2. **k-nearest neighbor graphs.** In this way, the goal is to connect vertex v_i with vertex v_j if v_j is among the k nearest neighbors of v_i. This way is usually used in image segmentation where each pixel has its neighbourhood defined by 4 or 8 pixels.

3. **The fully connected graph.** This way means connecting all points the pairwise similarity of which is positive. This construction is usually chosen if the similarity function itself already encodes mainly local neighborhoods. An example of such a similarity function is the Gaussian similarity function ($s(x_i,x_j) = e^{-\frac{||x_i-x_j||^2}{2\sigma^2}}$). The parameter σ controls the width of the neighbourhood.

Spectral clustering can be performed by three different algorithms [19]:

1. Unnormalized spectral clustering

2. Normalized spectral clustering according to Shi and Malik

3. Normalized spectral clustering according to Ng, Jordan, and Weiss

4 Proposed approach

This article focuses on behavioral modeling and analysis of animals that live in a closed environment. This simplifies the abstraction of the behavior since the number of possible actions of animals in a closed environment is limited. The proposed approach has been illustrated on a herd of cows containing one hundred of individuals.

The transition and emission matrices of the hidden Markov model are estimated using a sequence of states and sequence of observable values, therefore a set of possible states and a set of possible observable variable values have to be predefined. Each element of the sequence describes one minute of animal's behavior.

Five possible states that represent actions which an animal can do are defined. These states are denoted $S1, S2, S3, S4, S5$. An animal is considered to be in the state $S5$ if it is not in any of the states $S1$-$S4$. States S1, S2, S3 and S4 correspond to eating, drinking, resting and being milked.

Two observable variables are calculated for each state except the state $S5$:

1. the duration of the last occurrence of the state,

2. the time elapsed since the last occurrence of the state.

These eight variables are discretized using binning by equal frequency into four or five bins. The ninth observable variable represents a daytime which is discretized into six blocks of four hours. The Cartesian product of the value sets of observable variables contains, after discretization, 1,500,000 combinations of values.

To avoid zero probabilities in the emission and transition matrices, empirical transition and emission matrices averaged over all models of respective animal are used as initial estimates. Zero probabilities in those initial estimates are replaced with small probabilities estimated as one over the number of elements in the matrix.

Since the assumption of normally distributed observations is in our case not valid, we use an approximation of the Kullback-Leibler divergence, recalled in Subsection 3.2, as the similarity measure for HMM clustering. Using this approximation, a distance matrix D for the set of HMMs constructed for the individual anamals is computed, i.e., $D_{i,j}$ represents the KL divergence of the j-th from the i-th most likely sequence ($D_{i,j} = d_{\mathrm{Vit}}(\lambda_i, \lambda_j)$). The properties of the KL divergence imply that the matrix is real valued, non-symmetric and its diagonal has zero values.

Hierarchical and spectral clustering were applied to cluster the models. A parameter of hierarchical clustering is the kind of linkage. Since HMMs are not points in an Euclidean space, the single, average or complete linkage can be used. An optimal number of clusters can be determined by the domain expert from the dendogram of the hierarchical clustering.

For the spectral clustering, a fully connected graph based on the Gaussian similarity function is used as the similarity graph. Estimation of the parameter sigma of the similarity function is based on the optimization of balance of clusters.

The results of clustering were subsequently analysed using descriptive characteristics (e.g., age and weight of the animal) that were not among the observable variables.

5 Experimental results

The experiment is devided into two parts, behavioral modeling and clustering of the models. Section 5.1 describes results of the modeling using HMMs and Section 5.2 presents the results of clustering.

5.1 Hidden Markov modeling

The dataset that consists of ten consecutive days was split into train and test sets in a ratio 9:1. Parameters of a model were estimated with data of nine consecutive days and the model was evaluated with data of the tenth day. For each

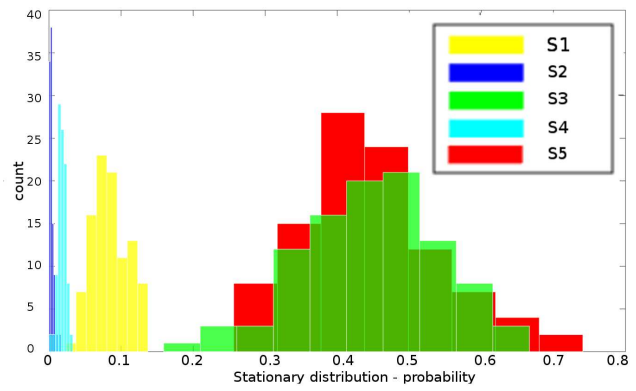

Figure 2: For each animal a stationary distribution of its Markov chain is calculated from a sequence of 14400 measurements. Values of these distributions are visualized for each state separately in a distribution.

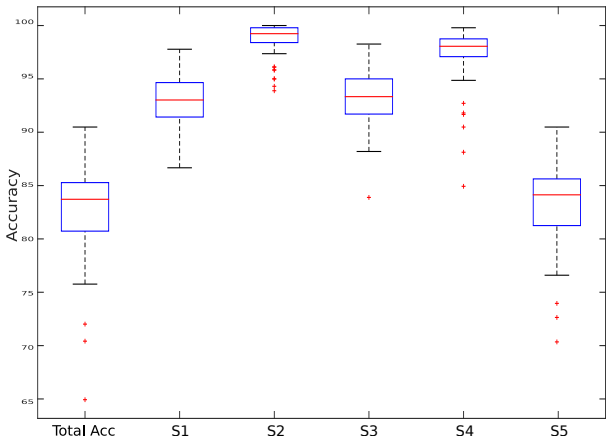

Figure 3: Accuracy of Viterbi's sequence.

animal, transitions and emissions matrices were estimated using the Baum–Welch algorithm.

Hidden Markov models were evaluated in two different ways, visually by a domain expert and using the Viterbi's sequence. The domain expert checks if the real animals' behavior is consistent with the stationary distribution of the resulting Markov chain, which describes the behavior to which the Markov chain converges. Consequently, the stationary distribution describes how much time an animal spends doing some activity in comparison with all activities.

Stationary distributions were computed for each animal separately. A histogram of the probabilities of individual states in the stationary distributions is shown in Figure 2.

Probabilities of the stationary distributions averaged over all animals are in Table 1.

State	Average probability
S1	8.5%
S2	0.4%
S3	44.8%
S4	1.8%
S5	44.5%

Table 1: Average stationary distribution.

The second way of evaluation is using the Viterbi's sequence. The Viterbi's sequence determines the most likely sequence of states given a sequence of observations and is denoted $v_1 v_2 .... v_T$. To get an accuracy of the prediction of a sequence of states, the real sequences of states are element-wise compared with the Viterbi's sequences. The accuracy is calculated as number of elements that do not differ divided by length of sequence. It was calculated separately for each state as well as for whole sequence.

The total accuracy was calculated as

$$\mathrm{Acc} = \frac{|\{i : v_i = q_i\}|}{T} \tag{15}$$

and an accuracy of a state s was calculated as

$$\mathrm{Acc}_s = \frac{|\{i : v_i = q_i, q_i = s\}|}{|\{j : q_j = s\}|}. \tag{16}$$

Resulting models are able to predict animals' actions with an average accuracy of 83.86%. It can be seen that the state S5 has the worst accuracy from all states. It is caused by a delay. A model can relatively precisely detect that an action has started but the detection of a termination of the action is delayed. It causes that the state S5 has worse accuracy than other states. Figure 3 visualizes the results of evaluation of the Viterbi's sequence.

5.2 Clustering

In this subsection, we discuss results and evaluation of the cluster analysis. The visualisation of a distance matrix is shown in Figure 4. A color of the point i,j represents distance between i-th animal's model and j-th animal's model. The lighter the color is the less similar the models are. The distance matrix is used by both clustering methods.

Linkages of hierarchical clustering can be visualized by dendrograms. Dendrogram is used to visually estimate an optimal number of clusters. The dendrogram of a complete linkage is shown in Figure 5. According to it, animals can be assigned into several approximately equally sized clusters.

The evaluation of clustering results is a difficult task since there are no references for validation. However, clusters can be validated by a domain expert based on descriptive characteristics related to animals (e.g., age, weight). These characteristics were not used for modeling, thus,

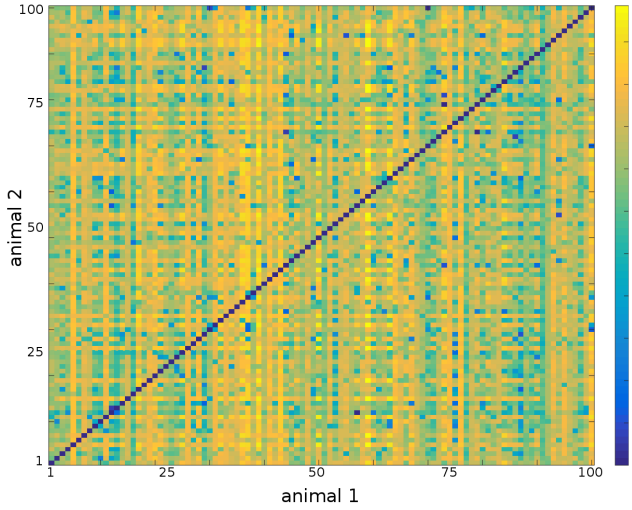

Figure 4: Kullback-Leibler divergence based distance matrix between estimated models. A color of the point i,j represents distance between i-th animal's model and j-th animal's model.

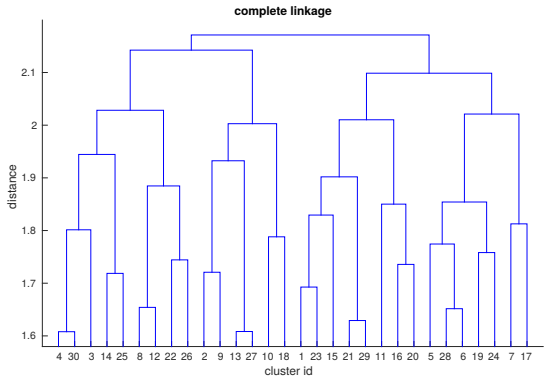

Figure 5: The dendrogram for a complete linkage.

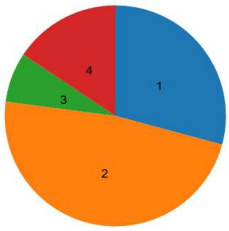

Figure 6: Proportions of four clusters, which were created using spectral clustering according to Shi and Malik.

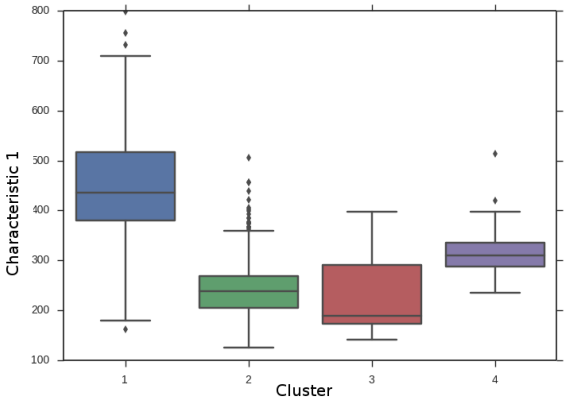

Figure 7: Boxplots of unobserved descriptive characteristic showing differences between four clusters.

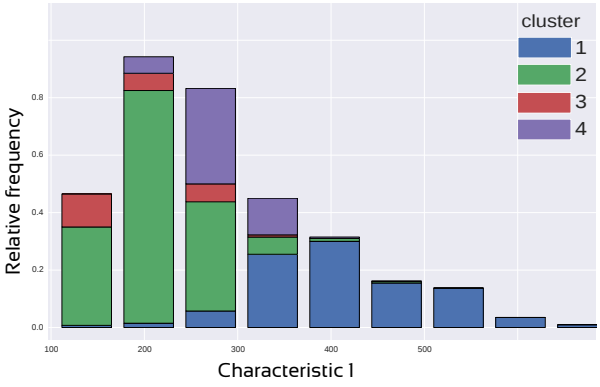

Figure 8: Distribution of unobserved descriptive characteristic showing differences between four clusters.

they can not influence the clustering. Figures 6, 7 and 8 show results of the spectral clustering according to Shi and Malik with a fully connected similarity graph with sigma equal to 0.65. In the figures can be seen that there are differences in descriptive statistics of such characteristic between individual clusters. For example, animals assigned to cluster 1 have significantly higher values of the characteristic than animals of other clusters. It indicates that the clustering is reasonable and meaningful.

6 Conclusion

Hidden Markov models are proved to be applicable to modeling of animal behavior represented as a sequence of states. According to their evaluation, the model is able to predict animals' actions with an average accuracy of 83.86%. Furthermore, the stationary distributions of the resulting models were validated by a domain expert. Cluster analysis was performed by classical clustering algorithms using the approximation of KL divergence. The clustering produces meaningful results and clusters animals into interpretable groups

Prediction can be further improved with better definition of states. In particular, the state S5 can be divided into more disjoint states. This may positively influence results of the clustering. Moreover, different modeling approaches commonly used for modeling sequences, such as dynamic Bayesian networks, conditional random fields, are intended to be researched, as well as their clustering possibilities.

References

[1] Jonathan Alon, Stan Sclaroff, George Kollios, and Vladimir Pavlovic. Discovering clusters in motion time-series data. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–375. IEEE, 2003.

[2] Longbing Cao and Philip S Yu. *Behavior computing*. Springer, London, 2012.

[3] Emanuele Coviello, Gert R Lanckriet, and Antoni B Chan. The variational hierarchical EM algorithm for clustering hidden Markov models. In *Advances in neural information processing systems*, pages 404–412, 2012.

[4] Qi Dai, Xiao-Qing Liu, Tian-Ming Wang, and Damir Vukicevic. Linear regression model of dna sequences and its application. *Journal of Computational Chemistry*, 28(8):1434–1445, 2007.

[5] Markus Falkhausen, Herbert Reininger, and Dietrich Wolf. Calculation of distance measures between hidden Markov models. In *In Proc. Eurospeech*, pages 1487–1490, 1995.

[6] Y. Guo, G. Poulton, P. Corke, G. J. Bishop-Hurley, T. Wark, and D. L. Swain. Using accelerometer, high sample rate GPS and magnetometer data to develop a cattle movement and behaviour model. *Ecological Modelling*, 220(17):2068–2075, 2009.

[7] John R Hershey and Peder A Olsen. Variational Bhattacharyya divergence for hidden Markov models. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4557–4560. IEEE, 2008.

[8] John R Hershey, Peder A Olsen, and Steven J Rennie. Variational Kullback-Leibler divergence for hidden Markov models. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 323–328. IEEE, 2007.

[9] Tony Jebara, Yingbo Song, and Kapil Thadani. Spectral clustering and embedding with hidden Markov models. In *Machine Learning: ECML 2007*, pages 164–175. Springer, 2007.

[10] Biing-Hwang Fred Juang and Lawrence R Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T technical journal*, 64(2):391–408, 1985.

[11] John L. Kelley. *General Topology (Graduate Texts in Mathematics)*. Springer, 1975.

[12] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 2001.

[13] Sang Wan Lee, Yong Soo Kim, and Zeungnam Bien. A nonsupervised learning framework of human behavior patterns based on sequential actions. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22(issue 4):479–492, 2010.

[14] Maja Matetić, Slobodan Ribarić, and Ivo Ipšić. Qualitative modelling and analysis of animal behaviour. *Applied Intelligence*, vol. 21(issue 1):25–44, 2004.

[15] Bill O'Brien, John Dooley, and Thomas J Brazil. Rf power amplifier behavioral modeling using a globally recurrent neural network. In *Microwave Symposium Digest, 2006. IEEE MTT-S International*, pages 1089–1092. IEEE, 2006.

[16] Vladimir Pavlović, James M Rehg, Tat-Jen Cham, and Kevin P Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 94–101. IEEE, 1999.

[17] Amy L. Sliva. *Scalable techniques for behavioral analysis and forecasting*. 2011.

[18] Padhraic Smyth et al. Clustering sequences with hidden Markov models. *Advances in neural information processing systems*, pages 648–654, 1997.

[19] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[20] Nelleke d. Weerd, Frank v. Langevelde, Herman v. Oeveren, Bart A. Nolet, Andrea Kölzsch, Herbert H. T. Prins, and W. F. Boer. Deriving animal behaviour from high-frequency GPS: Tracking cows in open and forested habitat. *PLoS One*, 10(6), 06 2015.

[21] Jie Yin and Qiang Yang. Integrating hidden Markov models and spectral analysis for sensory time series clustering. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[22] S.-X. Zhang. *Structured Support Vector Machines for Speech Recogntion*. PhD thesis, Cambridge University, March 2014.

[23] Shi Zhong. Semi-supervised sequence classification with HMMs. In Valerie Barr and Zdravko Markov, editors, *FLAIRS Conference*, pages 568–574. AAAI Press, 2004.

[24] Shi Zhong and Joydeep Ghosh. A unified framework for model-based clustering. *The Journal of Machine Learning Research*, 4:1001–1037, 2003.