SAND2018-2938C

Visualizing Clustering and Uncertainty Analysis of Multivariate Time-Series Data

Maximillian G. Chen, Kristin M. Divis, Laura A. McNamara, and J. Dan Morrow

Sandia National Laboratories, Albuquerque, NM

Sandia National Laboratories

Introduction

- Time-series data (panel data): multi-dimensional data where observations of multiple phenomena over multiple time periods are taken for the same subjects of interest
- Multivariate time-series data are intrinsic to the study of dynamic, naturalistic behavior.
- Probabilistic clustering models, such as the Hidden Markov Model(HMM), allow for identifying patterns in data under conditions of uncertainty.
- Most existing probabilistic clustering models, such as the Gaussian Mixture Model (GMM) assume observations are independent and identically distributed (i.i.d.), meaning these datasets have one observation for each subject.
- HMMs account for the temporal correlation between observations.

Motivation

Eye tracking Data:

- Eye trackers can generate voluminous spatio-temporal datasets comprising thousands of individual gaze samples that represent the calculated location of an individual's gaze against the display space.
- Gaze samples are aggregated using spatiotemporal thresholding algorithms into recognized behavioral indicators, such as saccades and fixations, that describe visual interaction with a stimulus.
- Current visualization tools are inadequate for assessing the performance of finite mixture models with eyetracking datasets, which are both spatially and temporally distributed.
- Question: **Can we use HMMs for probabilistic clustering of spatio-temporal eyetracking data?**

.

Eyetracking Dataset

- 16 human subjects
- Each subject looks at various points in an image, and the locations that the subject looks at are tracked in a one-hour long experiment in a constrained visual search task.
- A data point containing the spatial location of the subject's eye target is recorded every 17 milliseconds, so there are 25,000 sample points for the one subject throughout the four trials.
- See also Divis, Chen, McNamara, Morrow, & Perkins poster

.

Approach: Hidden Markov Model (HMM)

Model Assumptions:

- **Observed data**: m-variate time series of length T denoted by the general form
$$\mathbf{O}_{1:T} = (O_1^1, ..., O_1^m, O_2^1, ..., O_2^m, ..., O_T^1, ..., O_T^m).$$
- **Latent (hidden) states**: $\mathbf{S}_{1:T} = (S_1, ..., S_T)$
- **Model parameters**: θ
- **Covariates**: $\mathbf{z}_{1:T} = (\mathbf{z}_1, ..., \mathbf{z}_T)$

Visualization:

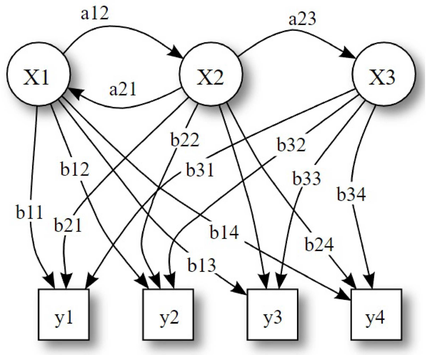

Figure: Probabilistic parameters of a HMM. X - hidden states; y - possible observations; a - transition probabilities; b - emission probabilities

Probabilities:

- $a_{ij}(\mathbf{z}_t) = P(S_{t+1} = j | S_t = i, \mathbf{z}_t)$: the probability of a transition from state i to state j with covariate $\mathbf{z}_t$. These are called the **transition probabilities**.
- $\mathbf{b}_{S_t}$: a vector of observation densities $b_j^k(\mathbf{z}_t) = P(O_t^k | S_t = j, \mathbf{z}_t)$ that provide the conditional densities of observations O_t^k associated with latent class/state j and covariate $\mathbf{z}_t$, $j = 1, ..., n$, $k = 1, ..., m$. These are called the **observation likelihoods** or **emission probabilities**.
- For full details on the HMM, see [1].

Uncertainty Quantification: We want to quantify the uncertainty of the predicted state of an observation at time t.

- **Posterior probability** of being in state j at time t given the observation sequence $\mathbf{O}_{1:T}$, covariates $\mathbf{z}_{1:T}$, and model parameters θ:
$$P(S_t = j | \mathbf{O}_{1:T}, \mathbf{z}_{1:T}, \theta'). \quad (1)$$
- **State Classification**: $S_t^* = \max_j P(S_t = j | \mathbf{O}_{1:T}, \mathbf{z}_{1:T}, \theta')$.
- **Classification Uncertainty**: $1 - \max_j P(S_t = j | \mathbf{O}_{1:T}, \mathbf{z}_{1:T}, \theta')$

Model Fitting and Selection:

- Use R package depmixS4 [2].
- Assume each individual input point (separate x- and y-coordinates of eye tracking data) follow a normal distribution.
- Select model using BIC criterion (lowest BIC value after fitting models with different number of hidden states).

References

- Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Proessing, Computational Lingistics, and Speech Recognition*, chapter 9: Hidden Markov Models, pages 122–141. Third edition draft edition.
- Ingmar Visser and Maarten Speekenbrink. depmixs4: An R package for hidden markov models. *Journal of Statistical Software, Articles*, 36(7):1–21, 2010.

Conclusions

- Create and demonstrate novel visualization methods for the clustering performance and assessing the clustering uncertainty.
- Demonstrate the validity of using the Hidden Markov Model with multivariate normal distribution assumption for clustering eyetracking data.
- Methods can be applied to time-series datasets in a wide array of application areas, such as radar and surveillance, medicine, and finance.

Future Work

- Fitting each (x,y) spatial data point as a multivariate normal distribution.
- Fitting different probability distributions (including nonparametric) and determining their goodness-of-fit relative to Gaussian distributions.
- Incorporating data from multiple trials, multiple tasks, and multiple subjects.

Application to Eye tracking Data

Plot of Eyetracking Data with Targets:

Figure: Plot of eyetracking data points for one subject, one task, and one trial. The points are assigned to 18 hidden states after fitting a HMM to the data. The percent signs represent the four targets on the image.

Uncertainty Plot:

Figure: Plot of eyetracking data points (for one subject, one task, and one trial) with clustering ellipsoids (95% level) and colored points representing the level of classification uncertainty.

Time Plot of Cluster Assignments:

Figure: Plot of cluster assignments of data points over time. With this plot, the chronological movement of the subject's eye with respect to the cluster number (and thus, location in the image) can be tracked.

Contact: mgchen@sandia.gov