Università di Pisa
Sant'Anna Scuola Universitaria Superiore Pisa
Consiglio Nazionale delle Ricerche
Università degli Studi di Cagliari

Book of Short Papers

SIS 2021

SIS 2021
PISA, 21-25 GIUGNO 2021

SIS Società Italiana di Statistica

Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

Contents

Analysis of clickstream data with mixture hidden markov models

Analisi dei clickstream data tramite i mixture hidden markov model

F. Urso, A. Abbruzzo, and M.F. Cracolici

Abstract Clickstream data is an important source of information for businesses, however it is not easy to manage this data and also to convert the information coming out from it in competitive advantage is not a trivial task. This study considers the application of mixture hidden Markov models to clickstream data extracted from a travel services company's e-commerce portal. We find clusters related to web users' browsing behaviour and geographical position that provide essential indications for developing new business strategies.

Abstract *I clickstream data sono un'importante fonte di informazioni per l'e-commerce, sebbene non siano semplici da gestire e convertire queste informazioni in un reale vantaggio competitivo non è un compito banale. In questo articolo, consideriamo l'applicazione dei mixture hidden Markov model a dati relativi al flusso di clickstream estratti dal portale e-commerce di un'azienda di servizi turistici. Sono stati individuati cluster relativi al comportamento di navigazione degli utenti e alla loro posizione geografica che forniscono indicazioni importanti per lo sviluppo di nuove strategie di business.*

Key words: Clickstream Data, Online browsing behaviour, Mixture hidden Markov models, Tourism 2.0, Web mining

---

F. Urso
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, e-mail: furio.urso@unipa.com

A. Abbruzzo
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, e-mail: antonino.abbruzzo@unipa.com

M.F. Cracolici
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, e-mail: mariafrancesca.cracolici@unipa.com

1 Introduction

Analyzing users' browsing behaviour when exploring e-commerce portals allows companies to gain significant advantages, such as the ability to classify potential customers, tailor their offers accordingly, or identify new opportunities that lead to changes in business strategies. Unfortunately, although the analysis of clickstream data provides essential information on the users' movements exploring a website (5; 3; 9; 6), it does not explain the underlying reasons behind their navigation choices. Furthermore, to develop effective marketing strategies, it would be desirable to identify users' subpopulations based on browsing behaviour. For this reason, statistical models with hidden variables such as mixture hidden Markov models (MHMMs) are a suitable tool for the analysis of clickstreams data. They allow to take into account two levels of uncertainty, a latent process whose evolution explains users' motivations to move from one page to another (13; 8; 7), and a hidden variable related to the presence of clusters representing browsing "profiles" (11; 15; 10). Here, we apply the MHMMs to data collected from the e-commerce portal of the PalermoTravel, a company operating in the hospitality sector, to analyze the differences in user behaviour by identifying the navigation profiles.[1] The paper is structured as follows: Section 2 illustrates the mixture of hidden Markov models. In Section 3, we have applied the model to identify browsing behaviour profiles taking into account user information such as geographic location obtained from IP addresses, access devices and access period.

2 Mixture Hidden Markov models

Hidden Markov models (2; 4; 14) allow analyzing time series whose evolution is supposed to depend on a latent Markov process. The mixture hidden Markov models enable to relax the hypothesis of a single population through latent variables that take into account the different longitudinal patterns in the sequences (15), identifying groups (clusters) of sequences assigned with specific probabilities derived from the data (11). This paper focuses on discrete MHMMs where the latent and the response are assumed discrete random variables.

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})$ be the generic i-th sequence of length T with card$|Y_i| = R$, $U_i = (U_{i1}, U_{i2}, \ldots, U_{iT})$ the i-th hidden random vector with card$|U_i| = S$ and assume n independent sequences. Let $M = \{M^1, M^2, \ldots, M^K\}$ be a set of HMMs, where $\Theta^k = \{\pi^k, A^k, B^k\}$ is the set of parameters for each sub-models M^k, related on each sub-population $k = 1, \ldots, K$. For each sequence Y_i, we define the prior cluster probabilities that the model parameters are the ones related to the k-th sub-model M^k as $P(M^k) = w_k$. The log-likelihood is computed as

[1] PalermoTravel is a pseudonym.

$$\ell(\Theta;Y) = \sum_{i=1}^{n} \log P(Y_i|\Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_{ik} \sum_{u} \pi_{u_1}^k b_{u_1}^k(y_{i1}) \prod_{t=2}^{T} a_{u_{t-1},u_t}^k b_{u_t}^k(y_{it}) \right), \quad (1)$$

where the hidden state sequences $u = (u_1, u_2, \ldots, u_T)$ take all possible combinations of values in the hidden state space S and where y_{it} are the observations of subject i at time t, $\pi_{u_1}^k = P(u_1 = s|\Theta^k)$ with $s \in \{1, \ldots, S^k\}$ is the initial probability of the hidden state at time $t = 1$ in sequence u for cluster k; $a_{u_{t-1},u_t}^k = P(u_t = j|u_{t-1} = i, \Theta^k)$ with $i, j \in \{1, \ldots, S\}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t in cluster k; and $b_{u_t}^k(y_{it}) = P(y_{it} = r|u_t = s, \Theta^k)$ with $s \in \{1, \ldots, S\}$ and $r \in \{1, \ldots, R\}$ is the probability that the hidden state of subject i at time t emits the observed state at t in cluster k. MHMM can be generalized to include time-constant covariates (15) that can be used to estimate cluster memberships w_{ik} of each sequence according to the following multinomial logistic model

$$w_{ik} = P(M^k|X_i) = \frac{e^{X_i\gamma_k}}{1 + \Sigma_{j=2}^{K} e^{X_i\gamma_j}}, \quad (2)$$

where γ_k is the set of coefficients associated with the vector of covariates X_i for observation i and the k-th class, and $\sum_{k=1}^{K} w_{ik} = 1$. The cluster posterior probabilities $P(M^k|Y_i, x_i)$ are obtained as

$$P(M^k|Y_i, X_i) = \frac{P(Y_i|M^k, X_i)P(M^k|X_i)}{P(Y_i|\Theta, X_i)}, \quad (3)$$

where $P(Y_i|\Theta, X_i)$ is the likelihood of the complete MHMM for subject i. In order to obtain the parameters estimates, the forward-backward algorithm (1; 12) can be used in MHMM context as illustrated by Vermunt et al. (15).

3 Results

MHMM has been used to analyze the difference in browsing behaviour among users of the PalermoTravel website. The data (log files) were collected in 2017 from September to December. They consist of 10,252 user sessions of maximum length T=20. These sessions are the sequences of pages viewed from the same IP address in a fixed period. Specifically, we do not consider the page names but their page category, corresponding to the thematic areas of the site: *Homepage*, *Attraction*, *Accommodation*, *Event*, *Experience*, *Service* and *Info* about the company. We consider three time-constant covariates collected from the website log files: IP address geographic area (i.e., Africa, Asia, East Europe, Italy, Latin America, Middle East, North America, North Europe, Oceania, Russia, South Europe), access device distinguishing PC and mobile and access month. These covariates are used to estimate prior cluster probabilities through the multinomial logistic model as in equation 2. Finally, using a selection procedure based on a combination of AIC and entropy, we

have selected the MHMM consists of 4 clusters and different hidden states in each cluster (i.e., 4,5,5,4) by applying a model selection procedure.

As an example, we show in Figure 1 a directed graph representing the path followed by users in cluster 1. Each pie graph represents a hidden state and edges are the transitions between states. Transition probabilities are displayed on the edges. The different colours and sizes of the pie slices represent emission probabilities of observed states (the pages' thematic area). The identified clusters are the following.

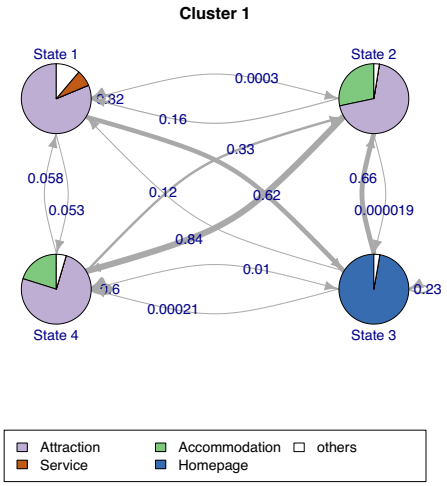

Fig. 1 Cluster 1 hidden Markov process structure. Vertexes represent hidden states, edges show the transition probabilities and the slices' color and size show emission probabilities. Emission probabilities lower than 0.05 are classified as "others".

Cluster 1 includes 15% of the sessions and has a high percentage of Northern Europeans who logged in the website from both PC and mobile, especially in October. Users start their session from state 3 (with probability 0.89). This state emits the observed state *Homepage* with probability 0.97. They move to state 2 with probability 0.66, then to state 4 with probability 0.84 and they stay in this state with probability 0.60. These states emit *Attraction* with probability 0.69 and 0.75 respectively. So, users appear to have a particular interest in general information about the region and less interested in the company's products, which is why the cluster was named **Information seeker**.

Cluster 2 (31.3% of the sessions) largely includes Italian, North American and Northern European users who explored the site using mostly their PC in November and December. Users start their session from state 3 with probability of 0.42 and stay in this state with probability 0.92. State 3 emits *Attraction* with probability 0.91. If users start from state 5 (with probability 0.23) they move to state 2 with probability 0.56 and stay in this state with probability 0.75. These two states emit *Event* with probabilities 0.72 and 0.62 respectively. If users start from state 1 with probability

0.19, they stay in this state with probability 0.96. State 3 emits *Accommodation* with probability 0.59. It seems that users primary interest remains viewing tourist attractions. Still, we also note that navigation can view seasonal events (a sign of the desire to select a period of visit) and apartments. This cluster was named **Potential tourist**.

Cluster 3 (41.3% of the sessions) includes the majority of Italians, with accesses mainly from a PC and in September. Users start from state 2 with probability 0.56. This state emits *Homepage* with probability 0.98. Then they move to other states with probabilities not too different. If they start from state 4 with probability 0.25, they will stay in this state with probability 0.95. This state emits *Accommodation* with probability 0.93. Users are interested in viewing and comparing tourism products, with less interest in the information pages presumably having prior knowledge of the Sicily region, so, this cluster was named **Expert tourist**.

Cluster 4 (12.4% of the sessions) includes most Asians and East Europeans and is characterized by the lowest percentage of access via mobile. Users start from state 3 with probability 0.82 and stay in this state with probability 0.89. State 3 emits *Homepage* with probability 0.99. So, they focused only on viewing the home page and move to different areas of the site. This cluster contains an interesting subgroup of users who would stay there if it reached state 2, this state emits *Info* with probability 0.99. This interest in information relating to the tourism company could be attributed to companies interested in partnership relationships. In light of these considerations, this cluster was named **Casual explorer or Potential partner**.

In summary, focusing on the first three profiles as a representation of the user's interest in purchasing a holiday package, we note that most users are categorized in profiles relating to medium and high interest: cluster two (Potential tourist) and cluster three (Expert tourist). Italian users are mostly present in group 3 and scarcely present in the information seeker profile. Users viewing both tourist information and products (cluster two) are mainly North Americans. In contrast, most North European countries are distributed in all first three profiles, making up most profile 1 of information seeker related to a lack of interest in the company's products. Regarding cluster 4, the users are from countries with a greater "cultural distance" from the Italian such as Slavic or Asian countries. Although scarcely present in the sample, these users are all in this profile showing a superficial interest (rarely accessing areas of the site other than the Homepage) or not purchasing oriented (e.g. potential partners). These results highlight that two users' target exploring the website exist, which come out out two different business models i.e. the business-to-consumer (already adopted by the analyzed firm), and the business-to-business.

In light of the above results, the company should consider making the website more attractive to potential customers from non-Western countries and consider selling products and services to other companies.

References

[1] Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state markov chains. The annals of mathematical statistics **37**(6), 1554–1563 (1966)

[2] Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The annals of mathematical statistics **41**(1), 164–171 (1970)

[3] Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S.: Visualization of navigation patterns on a web site using model-based clustering. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 280–284 (2000)

[4] Collins, L.M., Wugalter, S.E.: Latent class models for stage-sequential dynamic latent variables. Multivariate Behavioral Research **27**(1), 131–157 (1992)

[5] Cooley, R.W., Srivastava, J.: Web usage mining: discovery and application of interesting patterns from web data. Citeseer (2000)

[6] Das, R., Turkoglu, I.: Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Expert Systems with Applications **36**(3), 6635–6644 (2009)

[7] De Angelis, L., Dias, J.G.: Mining categorical sequences from data using a hybrid clustering method. European Journal of Operational Research **234**(3), 720–730 (2014)

[8] Dias, J.G., Vermunt, J.K.: Latent class modeling of website users' search patterns: Implications for online market segmentation. Journal of Retailing and Consumer Services **14**(6), 359–368 (2007)

[9] Eirinaki, M., Vazirgiannis, M., Kapogiannis, D.: Web path recommendations based on page ranking and markov models. In: Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 2–9 (2005)

[10] Helske, S., Helske, J.: Mixture hidden markov models for sequence data: The seqhmm package in r. arXiv preprint arXiv:1704.00543 (2017)

[11] Van de Pol, F., Langeheine, R.: Mixed markov latent class models. Sociological methodology pp. 213–247 (1990)

[12] Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)

[13] Scott, S.L., Hann, I.H.: A nested hidden markov model for internet browsing behavior. Marshall School of Business pp. 1–26 (2006)

[14] Vermunt, J.K., Langeheine, R., Bockenholt, U.: Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. Journal of Educational and Behavioral Statistics **24**(2), 179–207 (1999)

[15] Vermunt, J.K., Tran, B., Magidson, J.: Latent class models in longitudinal research. Handbook of longitudinal research: Design, measurement, and analysis pp. 373–385 (2008)