Co-clustering TripAdvisor data for personalized recommendations

Co-clustering di dati TripAdvisor per un sistema di raccomandazioni personalizzato

Giulia Pascali, Alessandro Casa and Giovanna Menardi

Abstract TripAdvisor is one of the largest travel websites. Among the provided services, it aids users with suggestions about attractions, accommodations, restaurants, etc., based on a wide system of reviews. In fact, users looking for suggestions shall sort through opinions posted by any kind of other users, possibly with different preferences and travelling behaviour. The aim of this work is to provide a personalized recommendation system to integrate the TripAdvisor services, based on the identification of similar travel products rated by similar users. Some alternative models of co-clustering are considered, to handle user rates in the form of ordinal data, and to account for missing values, due to the intrinsic fact that each user rates only a small subset of the considered travel products. Possible extensions are discussed to include additional information in the model, based on products and users characteristics.

Abstract *TripAdvisor è una delle più grandi piattaforme web dedicate al turismo. Tra i servizi offerti, il più noto è quello di fornire consigli di viaggio basati sulla valutazione delle attrazioni turistiche, hotel, e ristoranti, da parte degli utenti. Tuttavia, i viaggiatori alla ricerca di suggerimenti devono districarsi tra opinioni fornite da ogni tipologia di utente, con preferenze ed abitudini di viaggio anche molto diverse. L'obiettivo di questo lavoro è quello di fornire un sistema di raccomandazioni personalizzato ad integrazione di quello fornito da TripAdvisor. Si considerano alcuni modelli di co-clustering volti a gestire le valutazioni degli utenti nella forma di dati ordinali tenendo conto al tempo stesso della presenza di dati mancanti, problema intrinseco legato alla tendenza del singolo utente a recensire solo un esiguo sottoinsieme di prodotti turistici tra quelli disponibili. Vengono inoltre discusse possibili estensioni dei modelli, finalizzate a includere ulteriori informazioni riguardo alle caratteristiche dei prodotti recensiti e degli utenti.*

Key words: co-clustering, ordinal data, recommendation systems

Giulia Pascali, Alessandro Casa, Giovanna Menardi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova
via C. Battisti 241, 35121, Padova; e-mail: giulia.pascali@studenti.unipd.it, casa@stat.unipd.it,menardi@stat.unipd.it

1 Introduction

TripAdvisor[1] is one of the largest travel websites, providing users with suggestions about trip plans, attractions, accommodations, restaurants etc. While the main tool for advising is a wide system of reviews, users looking for suggestions shall sort through opinions provided by any kind of other users, possibly with different preferences and travelling behaviour. In this immense set of possible alternatives, personalized recommendations can come to the user aid and help addressing a choice.

Recommendation systems are becoming increasingly sophisticated in the e-commerce era aiming at providing users with more and more targeted and personalized advices. Especially based on machine learning techniques [5], recommendation systems establish connections among users and items based on the preferences of similar users or on product similarity, in order to predict users' choices. A statistical approach which lends itself to the purpose of creating personalized recommendations is known as *co-clustering*. Aimed at jointly identifying clusters of observations and variables i.e., in the specific framework, users and items, co-clustering can be seen as a way to combine recommendation systems based on collaborative filtering - that account for similarities among users only - and content-based techniques, considering similarities among items.

In this work we aim at providing a recommendation system to be integrated within the TripAdvisor services. Some alternative models of co-clustering are considered, to handle user rates in the form of ordinal data, and to account for the intrinsic problem of missing values, as each product is rated by a small subset of users only. Possible model extensions are discussed to include additional information in the model, e.g. based on product tags or user reviews.

After describing the data and their main features (Section 2), we introduce the Latent Block Model to perform co-clustering (Section 3) and overview some possible specifications for the problem at hand. Results on their application on TripAdvisor data are illustrated and discussed (Section 4), along with some model extensions.

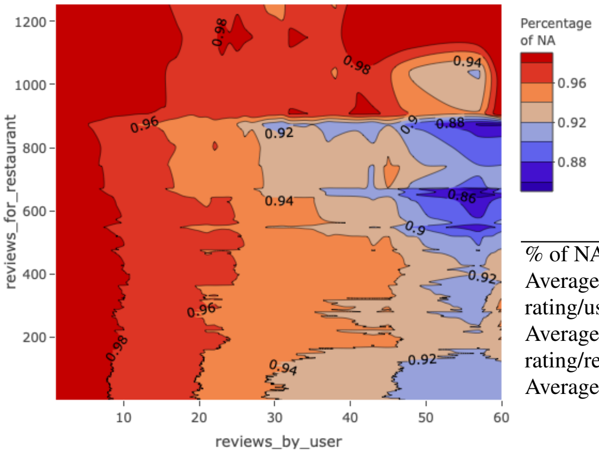
2 Data Description

Data at hand have been downloaded from the web[2] via web scraping, and refer to all the restaurants and bars located in the province of Padova and rated on TripAdvisor. The considered time horizon spans from the 1st of August 2011 (date of the first review) to the 15th of November 2018 (date of the download).

The original data set includes 709 restaurants and 42.263 users. Each rate consists of an integer number between 1 (terrible) to 5 (excellent). A whole amount of 97555 ratings is observed, leading to a very sparse (99.7% of missing values) utility

[1] `www.tripadvisor.com`

[2] `https://www.tripadvisor.it/Restaurants-g187867-Padua_Province_of_Padua_Veneto.html`



	Complete data	Reduced data
% of NAs	99.7	94
Average # of rating/user	1.72	26
Average # of rating/restaurant	102.3	2.34
Average rating	3.94	3.62

Fig. 1: Left: empirical distribution of the missing values as the number of reviews per user and the number of reviews per restaurant varies. Right: descriptive statistics of complete and reduced data.

matrix - i.e. a matrix with cells reporting the ratings and rows and columns are associated with users and restaurants respectively.While one specific strategy is required to handle such a large amount of missing values, we have proceeded with a preliminary reduction of the data in order to limit the problem as far as possible. Fig. 1 highlights how the percentage of missing data changes when removing a number of reviewing users and restaurants being reviewed. Hence, to give more weight to the most informative part of the data, a reduced utility matrix with 94% of missing data has been extracted from the original one for the analysis, including 500 restaurants (each of them rated at least by 10 users) and 45 users (each of them having rated at least 25 restaurants).

Additional information for each rate may (but not necessarily does) include the price range, the total score, a tag indicating the type of cuisine of the restaurant (e.g. italian, vegetarian, japanese etc...), and a text review.

3 Co-clustering for ordinal data

In order to build a personalized recommendation system, based on both user and item similarities, a suitable framework is represented by the so-called *co-clustering* approach. Co-clustering aims at providing a joint partition of rows and columns of a data matrix. Several approaches have been proposed in literature, following either heuristic or probabilistic perspectives [3]. Among the latter ones, the most considered one is, unarguably, the *latent block model* (LBM).

Let $\mathbf{x} = (x_{ij})_{1<i<n,1<j<p}$ be the data at hand, where, in our case, each $x_{ij} \in \{1,\dots,M\}$ corresponds to the (possibly missing) rate given by the *i-th* user to the *j-th* restaurant. Two Multinomial latent variables $\mathbf{z} = (z_{ik})_{1<i<n,1<k<K}$ and $\mathbf{w} = (w_{jl})_{1<j<p,1<l<L}$ are introduced to describe respectively the row and the column cluster membership, where $z_{ik} = 1$ if observation i belong to row-cluster k and the same holds for w_{jl}, and K, L the number of row and column clusters. In the LBM framework the independence between $\mathbf{z}$ and $\mathbf{w}$ is assumed and, conditionally on $\mathbf{z}$ and $\mathbf{w}$, the $n \times p$ observations x_{ij} are also independent. A general latent block model for $\mathbf{x}$ is specified as follows:

$$p(\mathbf{x};\theta) = \sum_{z \in Z} \sum_{w \in W} p(\mathbf{z};\theta) p(\mathbf{w};\theta) p(\mathbf{x}|\mathbf{z},\mathbf{w};\theta) \ , \qquad (1)$$

where Z and W are respectively the rows and columns partitions, $p(\mathbf{z};\theta) = \prod_{ik} \rho_k^{z_{ik}}$ and $p(\mathbf{w};\theta) = \prod_{jl} \delta_l^{w_{jl}}$; $p(\mathbf{x}|\mathbf{z},\mathbf{w};\theta) = \prod_{ijkl} p(x_{ij};\alpha_{kl})^{z_{ik}w_{jl}}$ and $\theta = (\rho_k, \delta_l, \alpha_{kl})$.

Coherently with the nature of the data, we need to assume for $p(\cdot;\alpha_{kl})$ a probabilistic model which handles ordinal data. We focus on the proposal by Jacques and Biernacki [4] and on the work by Corneli et al. [2].

The former approach specifies $p(\cdot;\alpha_{kl})$ as a BOS model [1], a probability distribution designed for ordinal data and governed by the parameter $\alpha_{kl} = (\mu_{kl}, \tau_{kl})$, $\mu_{kl} \in \{1,...,M\}, \tau_{kl} \in [0,1]$, with, respectively, the roles of location and precision.

In the latter approach [2], the generative model is based on a Gaussian latent random variable with parameters $\alpha_{kl} = (\mu_{kl}, \sigma_{kl}^2)$ sharing the same rationale as the cumulative probit model, i.e. the ordinal categories are defined by continuous intervals between pre-determined cut-points $\gamma_1, \dots, \gamma_{M-1}$. The model can manage missing data possibly not at random, via the introduction of a Bernoulli variable A_{ij} with parameter π_{kl}, describing the presence of a cell observation. Hence,

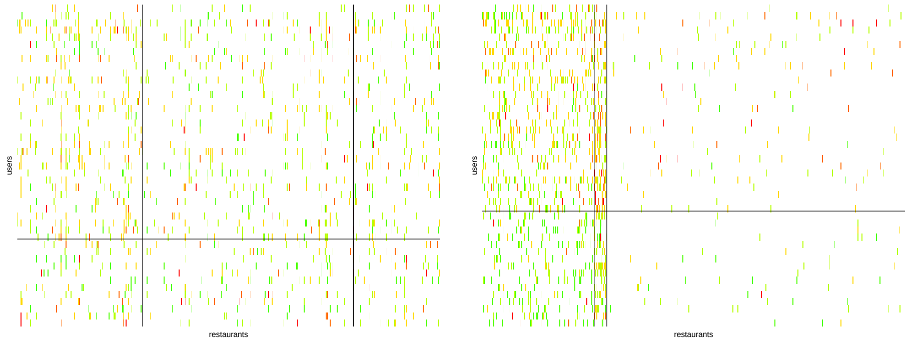
$$p(X = m;\alpha_{kl}|A_{ij} = 1) = \Phi\left(\frac{\gamma_m - \mu_{kl}}{\sigma_{kl}}\right) - \Phi\left(\frac{\gamma_{m-1} - \mu_{kl}}{\sigma_{kl}}\right) \qquad (2)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a Gaussian variable.
In both the models, parameters are estimated via maximum likelihood.

The reader may refer to the original works for further details.

4 Results and discussion

In the following, we analyze and compare the results of the application of the considered methods on the data introduced in Section 2. The number of co-clusters has been determined via the optimization of an information-based criterion when resorting to the approach by [2], and set to the same number when considering the work by [4], since the lack of a way for handling missing values discourages the automatic selection.



(a) Method of Jacques and Biernacki [4]

(b) Method of Corneli et al [2]

Fig. 2: Utility matrices of TripAdvisor Padova data reorganized according to the results obtained. An empty cell corresponds to a restaurant that has not been rated by the user. Different rates are associated to different colors, ranging from negative reviews (red) to positive ones (green).

In Fig. 2 the utility matrices reorganized according to the results obtained with the two approaches are illustrated. From a graphical inspection it stands out a broadly uniform distribution among the blocks obtained via the application of [4]; the intuition is confirmed by the analysis of the estimated parameters of the model. As a possible responsible for this behaviour which jeopardizes an effective interpretation of the results, we can identify the *missing at random* assumption inducing comparable frequencies of missing values in each co-cluster. Conversely, a joint inspection of Fig. 2 and Fig. 3 allows to obtain some interesting insights on the data when the method [2] is applied. Generally, it can be noticed that one row cluster groups users assigning lower rates with respect to the users in the other row cluster. The percentage of missing values varies among blocks; two co-clusters represent rarely reviewed restaurants, with average and respectively good ratings. Two further column cluster represents groups of largely reviewed restaurants, with row rates distributed coherently with the considerations above. Thus, the method [2] appears more appropriate than [4], both for the way it handles missing observations and for computational reasons.

Cluster homogeneity suggests a non-negligible informativeness of the customer-oriented recommendation system driven by the results. For example it would be reasonable to suggest, to customer in the second row cluster, the restaurants in the second column cluster where they do not have already eaten. It seems reasonable to inversely weight the suggestions with the percentage of missing values in each co-cluster, according to the rationale for which recommendations for rarely reviewed restaurants would have a greater degree of uncertainty.

Results suggest further room for improvement of the model [2], thanks to the availability of additional information about restaurants and users characteristics, such as the price range and type of food sold. The mean of the underlying latent Gaussian random variable can be modeled via a regression function, depending on row and column specific features. This would allow us to obtain covariate-dependent block-specific means and to better characterize the induced co-clusters. As a by product, it would be possible to evaluate if covariates show a different impact on the response variable, depending on the specific block. Despite naturally sensible, the idea requires non-trivial modifications of the estimation procedure which are left for future research.

References

1. Biernacki, C., Jacques, J. (2016) Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing* 26 (5), 929–943.
2. Corneli M, Bouveyron C. and Latouche P. (2019) Co-Clustering of ordinal data via latent continuous random variables and a classification EM algorithm. *https://hal.archives-ouvertes.fr/hal-01978174*. HAL Id: hal-01978174
3. Govaert G., Nadif M. (2013) Co-Clustering, Wiley-IEEE Press.
4. Jacques, J. and Biernacki, C. (2018) Model-based co-clustering for ordinal data. *Computational Statistics & Data Analysis*, 123, 101-115.
5. Ricci, F, Lior R., and Bracha S. (2015) *Recommender Systems Handbook*. Springer.

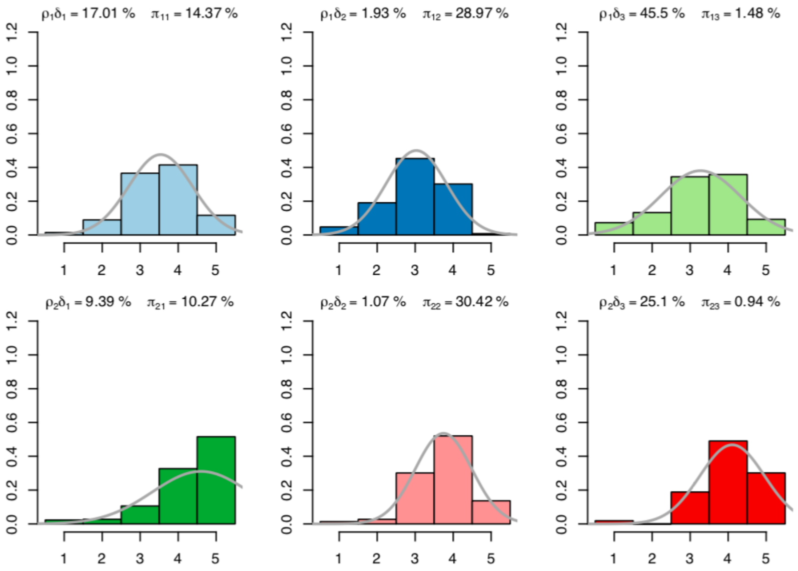


Fig. 3: Histograms of rates of the co-clusters identified by [2] with superimposed the estimated latent Gaussian variable. The estimated parameters and percentages of missing values are also reported.