ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

The role of feature selection in personalized recommender systems

Roger Bagué-Masanés[1], Verónica Bolón-Canedo[1] and Beatriz Remeseiro[2] *

1- CITIC, Universidade da Coruña, A Coruña, Spain

2- Universidad de Oviedo, Gijón, Spain

Abstract. Recommender systems suggest products to users, based on their popularity or the users' preferences. This paper proposes a hybrid personalized recommender system based on users' tastes and also on information available about items. We used a dataset downloaded from TripAdvisor, which contains some information from restaurants (items), such as price range or special diets. Feature selection techniques are employed to analyze the impact that each variable has on personalized recommendations, allowing us to understand not only the process underlying the recommendation to favor the transparency of the system, but also what users value the most when choosing a restaurant.

1 Introduction

Recommender systems (RS) are techniques that provide suggestions for items that may be helpful or interesting to users [1]. These suggestions are related to various decision-making processes, e.g., what product to buy, what music to listen to, or what restaurant to eat at. The need to provide users with personalized suggestions is not only a desirable feature, but with the large amount of data that they handle and offer, it has become a necessity.

There are several types or categories of RS, depending on the information they use. On the one hand, there are personalized RS, which suggest articles or services based on user preferences; and non-personalized RS, which are based on the popularity of the items. On the other hand, collaborative RS leverage other users' ratings to provide the recommendations, whilst content-based RS use descriptive keywords associated with the items. More advanced RS use not only users' ratings, but also their reviews and/or the images they took of the items. Finally, it is worth mentioning that most RS used in practical applications are a combination of these basic models.

In this context, we propose to take into account the users' ratings along with a set of details of the items, which in our case of study are restaurants found on TripAdvisor[1]. TripAdvisor is an online platform that recognizes millions of opinions about certain businesses in the tourism industry (e.g., restaurants, hotels, bars, nightclubs, cruises). In the case of restaurants, which are the focus

*This work has been supported by the Spanish Ministerio de Economía y Competitividad (Grant PID2019-109238GB, subprojects C21 and C22) and by the Xunta de Galicia (Grants ED431C 2018/34 and ED431G 2019/01) with part of European Union ERDF funds.

[1] https://www.tripadvisor.com/

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

on this work, there are a total of 208 different details about them, within the following categories: *price range*, *cuisine*, *special diets*, *meals*, and *features*.

This large amount of data can be a problem when it comes to algorithm training. To face this problem, we propose to use feature selection (FS) techniques to discover the important features and discard the irrelevant ones. The correct selection of features can lead to enhancement of the inductive learner, whether in terms of learning speed, generalization ability, or simplicity of the induced model. In addition to this, the fact of having a smaller subset of features has the potential of reducing measurement costs, and allowing a better understanding of the domain [2]. In particular, we will focus on filter methods, as they have a low computational cost and are suitable for high-dimensional data.

The use of FS in RS has not been broadly explored, apart from the work of Afoudi et al. [3], who compared different FS techniques to improve performance in RS. Their proposal is focused on content-based systems and presents an RS that classifies restaurants by priority of order. Cataltepe et al. [4] performed FS on each user's profile to make predictions about movie ratings. Unlike the previous works, we will analyze the impact of the FS when it comes to making the recommendation more understandable and knowing what users take into account when choosing a restaurant.

The goal of this paper is to analyze the impact of applying FS methods to the available information of restaurants (e.g., *price range*, types of *cuisines*, etc.) in the context of personalized recommendations. In particular, the main contributions of this research are three-fold: (1) a hybrid personalized RS for restaurants that benefits from users' opinions, as collaborative RS do, and from restaurant details, as content-based RS do; (2) an analysis of the impact of different restaurant details (e.g., *price range*, types of *cuisines*, etc.) in the context of personalized recommendations; and (3) a study of the role of FS not only to provide a recommendation as good as the one obtained with all the restaurant details but using a smaller subset of them, but also to understand the process underlying the recommendation to favor the transparency of the system.

2 Materials and methods

2.1 Dataset and data preparation

We used a modified version of the dataset proposed in [5], which contains the reviews from restaurants downloaded from TripAdvisor of the cities of Gijón, Barcelona, and Madrid[2]. The information available for each review is: an identifier of the restaurant (*restaurantId*), an identifier of the user who wrote the review about the restaurant (*userId*), and a number that represents the score (from 1 to 5 stars) given by the user to the restaurant (*rating*).

The original dataset was expanded to include the details of the restaurants, which were downloaded with the Scrapy[3] framework. Table 1 shows some figures

[2] The dataset is available in: `https://doi.org/10.5281/zenodo.5644892`

[3] `https://scrapy.org/`

	Gijón	Barcelona	Madrid
Users	24900	170974	230707
Restaurants	529	5294	6146
Restaurants w/o details	14	430	279
Reviews (% likes)	48108 (84.1%)	404946 (86.9%)	561587 (86.3%)

Table 1: Information available for each city.

related to our dataset in terms of users, restaurants, and reviews. The details of the restaurants are optional and can be grouped into five categories: (1) *price range*, with the lowest and the highest prices; (2) *cuisines*, a multi-label attribute with a maximum of 5 labels out of a total of 156 different options; (3) *special diets*, a multi-label attribute with a maximum of 5 labels out of a total of 5 different options; (4) *meals*, a multi-label attribute with a maximum of 6 labels out of a total of 6 different options; and (5) *features*, a multi-label attribute with a maximum of 39 labels out of a total of 39 different options. Regarding the ratings, we consider that users like a restaurant if they have rated it with at least 3 stars and dislike it otherwise.

As part of the restaurant information is optional, in some categories there is an important amount of missing data (e.g., 59%-77% in *price range*, depending on the city). In the presence of missing values, we decided to impute the missing data with zero to train the system.

The next step was to determine how to represent the restaurant details as a vector. In the case of the *price range*, it was divided into two values that correspond to the lowest and highest prices, respectively. As the rest of categories are multi-label, with N possible values, we converted each one into N binary characteristics. After this pre-processing, the dataset contains 208 different binary attributes corresponding to all the restaurant details. Finally, each *restaurantId* was encoded as a number from 1 to M, where M is the number of restaurants. The same procedure was applied to encode the *userId*.

After the pre-processing step previously described, the dataset was divided into different partitions used for training and evaluation purposes. Given the nature of the dataset, standard divisions with a percentage of samples per partition are not an option. For this reason, the partitions were created through an ad-hoc procedure: (1) among all the positive reviews of each user, one of them goes to the test set and the rest to the training set; (2) the previous step is applied to the negative reviews; and (3) if there is a user and/or restaurant in the test set that does not appear in the training set, the corresponding review is moved to the training set. Finally, the same procedure was applied to the obtained training set to divide it again into training and validation sets.

As this dataset is highly unbalanced in favor of the *like* class, negative sampling [6] is used to balance the training dataset by creating negative samples, thus having the same or similar number as the positive ones. In this research, the procedure consists in iterating over all users and checking their number of reviews. If a user has more positive reviews than negative reviews, the difference between them is calculated to determine the number of negative samples to generate. Those negative reviews are randomly created by choosing restaurants that the user has never visited (assuming that the user does not like a restaurant never visited, which is a common practice in RS).

2.2 Learning methods

As FS method, we chose a very popular filter: mutual information maximization (MIM) [7]. This method ranks the features according to their importance with respect to the class using mutual information.

After this step, we used an RS to evaluate the impact of using FS on the available information on each restaurant. The target of the RS is to predict if a user likes or dislikes a certain restaurant. In terms of machine learning, it is a binary classification problem with the user and the restaurant details as inputs.

The proposed RS receives two input data: (1) a user, represented by a one-hot codification and mapped into a 64-dimensional vector; and (2) a restaurant, represented by means of their details and codified as a 64-dimensional vector through a fully connected (FC) layer followed by a rectified linear unit (ReLU) [8]. Next, a processing block (FC layer + ReLU + batch normalization [9]) is sequentially applied eight times. Finally, the last layer is composed of a FC layer with the sigmoid activation function, which outputs a probability. Note that the binary cross-entropy was used as the loss function and Adam [10] was selected as the optimization algorithm for stochastic gradient descent.

3 Experimental results

This section presents the results obtained by applying the MIM feature selection method to the TripAdvisor datasets, which are available for download[4]. The code to reproduce all the experiments is available in GitHub[5].

First, we used MIM to measure the impact of the different restaurant details on personalized recommendation. The top 20 details for each city can be seen on the supplementary material[6]. Focusing on these details, we can draw some interesting conclusions. Firstly, most of top 20 attributes belong to the category called *features*, followed by *special diets*. More specifically, the most important *features* are those related to payment methods and accessibility. Regarding *special diets*, they are selected in all cities, so they seem to be a decisive factor when choosing a restaurant (e.g., a vegetarian person would not go to a restaurant that only serves meat). Another important factor are *cuisines*. As the data are from Spanish cities, the most common labels include Mediterranean and Spanish.

It is worth noticing that 7 out of the 20 most important features are the same in the three cities analyzed, although they appear in different positions of the ranking. As can be seen on the supplementary material, the details shared

[4] The dataset is available in: `https://doi.org/10.5281/zenodo.6782602`

[5] `https://github.com/rbague5/TFM`

[6] `https://github.com/rbague5/TFM/blob/main/Supplementary%20Material.pdf`

between the three cities are: Gluten Free Options, Vegetarian Friendly, Vegan Options, min range, max range, Accepts Credit Cards, and Free WiFi. In short, what most people take into account when choosing a restaurant is the price and if they have any food needs (e.g., intolerance or allergy).

In order to check the impact of using a different number of details when feeding the proposed RS, we selected different values in the range [10, 50], according to the scores provided with MIM. We also trained the RS with all the available details (Baseline) to measure the impact of applying FS. Table 2 shows the results obtained for Gijón, Barcelona, and Madrid, which were evaluated in terms of true positive rate (TPR), true negative rate (TNR), area under the ROC curve (AUC), and balanced accuracy (BA).

Table 2: Summary of the recommender system results in the three cities, using MIM for feature selection (FS). Best results per city are in bold.

			FS - Number of details				
		Baseline	10	20	30	40	50
Gijón	TPR	0.6199	0.5703	0.7250	**0.7603**	0.6837	0.7373
	TNR	0.5957	**0.6409**	0.4554	0.4150	0.5125	0.4483
	AUC	**0.8279**	0.7912	0.8152	0.8160	0.8192	0.8033
	BA	**0.6078**	0.6056	0.5902	0.5876	0.5981	0.5928
Barcelona	TPR	0.7208	0.6607	0.7475	**0.7724**	0.7119	0.7389
	TNR	0.4175	**0.4574**	0.3815	0.3507	0.3901	0.3396
	AUC	0.8046	0.7748	0.7717	0.7969	0.8118	**0.8321**
	BA	**0.5692**	0.5591	0.5645	0.5616	0.5510	0.5393
Madrid	TPR	**0.7581**	0.5202	0.7404	0.7412	0.7107	0.7456
	TNR	0.4050	**0.5821**	0.4332	0.4365	0.4436	0.4392
	AUC	**0.8407**	0.7636	0.7978	0.8162	0.6073	0.6175
	BA	0.5816	0.5512	0.5868	0.5888	0.5772	**0.5924**

In general terms, all combinations produce very similar results with an AUC close to 80%. As can be seen, applying FS leads to an improvement of the detection of negative samples in comparison to the baseline. It is worth noting that using only the top 10 restaurant details achieves the best results in terms of TNR, for all three cities. Note that this means that this configuration is better than the others at detecting the minority class (negative reviews). This is particularly important in this context, as it is better to miss the recommendation of a restaurant that a user would like, than to recommend a restaurant that the user would not like. Furthermore, the RS obtains these results using only 10 of 208 available details, according to the MIM scores, thus generating a much simpler and easier interpretation of the model.

We can also see that, for two of the cities (Gijón and Barcelona), using only 30 details is enough to obtain the best results in terms of TPR. When focusing on AUC, applying FS achieves the best results in Barcelona (50 details); and in terms of BA, in Madrid (50 details).

4 Conclusions

RS often have to deal with large datasets, both in the number of features and samples. For this reason, we propose the use of FS to obtain simpler and more explainable models, without losing accuracy in the recommendation. In particular, we focus on evaluating the use of FS in the context of personalized recommendations. Firstly, we propose an RS for restaurants that, for each (user, restaurant) pair, predicts whether the user likes this restaurant or not. Users are represented by an artificial encoding (one-hot codification), whilst restaurants are represented by their details (e.g., *price range*, types of *cuisines*, etc.) Given that the number of details is very large, we used FS to analyze their relevance.

The experimentation carried out demonstrated the competitiveness of the prediction results, showing that using MIM for FS can improve the prediction of negative cases by 4.52% in Gijón, 3.99% in Barcelona, and 17.71% in Madrid compared to the Baseline, but using only 10 attributes, which means 4.81% of all available details. Analyzing the relevance of the different restaurant details, it is observed that the most important criteria when choosing a restaurant are: the type of special diets that it serves, the different forms of payment since more and more people pay digitally instead of in cash, and the accessibility for people.

References

[1] T. Mahmood and F. Ricci, “Improving recommender systems with adaptive conversational strategies,” in *20th ACM Conf. on Hypertext and Hypermedia*, 2009, pp. 73–82.

[2] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, “A review of feature selection methods on synthetic data,” *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, 2013.

[3] Y. Afoudi, M. Lazaar, and M. Al Achhab, “Impact of Feature selection on content-based recommendation system,” in *Int. Conf. on Wireless Technologies, Embedded and Intelligent Systems*, 2019, pp. 1–6.

[4] Z. Cataltepe, M. Uluyağmur, and E. Tayfur, “Feature selection for movie recommendation,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 24, no. 3, pp. 833–848, 2016.

[5] J. Díez, P. Pérez-Núnez, O. Luaces, B. Remeseiro, and A. Bahamonde, “Towards explainable personalized recommendations by learning from users’ photos,” *Information Sciences*, vol. 520, pp. 416–430, 2020.

[6] B. W. Yap, K. Abd Rani, H. A. Abd Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, “An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets,” in *Int. Conf. on Advanced Data and Information Engineering*, 2014, pp. 13–22.

[7] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[8] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” in *27th Int. Conf. on Machine Learning*, 2010.

[9] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Int. Conf. on Machine Learning*, 2015, pp. 448–456.

[10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations*, 2015, pp. 1–15.