A hybrid feature combination method that improves recommendations

Gharbi Alshammari[1], Stelios Kapetanakis[1], Abduallah Alshammari[1], Nikolaos Polatidis[1], and Miltos Petridis[2]

[1] School of Computing, Engineering and Mathematics
University of Brighton, Moulsecoomb
Campus, Lewes road, Brighton BN2 4GJ, UK
(g.alshammari, s.kapetanakis, a.alshammari1, n.Polatidis)@brighton.ac.uk
[2] Department of Computer Science
Middlesex University London,
The Burroughs, London NW4 4BT, UK
M.Petridis@mdx.ac.uk

Abstract. Recommender systems help users find relevant items efficiently based on their interests and historical interactions. They can also be beneficial to businesses by promoting the sale of products. Recommender systems can be modelled by applying different approaches, including collaborative filtering (CF), demographic filtering (DF), content-based filtering (CBF) and knowledge-based filtering (KBF). However, large amounts of data can produce recommendations that are limited in accuracy because of diversity and sparsity issues. In this paper, we propose a novel hybrid approach that combines user-user CF with the attributes of DF to indicate the nearest users, and compare the Random Forest classifier against the kNN classifier, developed through an investigation of ways to reduce the errors in rating predictions based on users past interactions. Our combined method leads to improved prediction accuracy in two different classification algorithms. The main goal of this paper is to identify the impact of DF on CF and compare the two classifiers. We apply a feature combination hybrid method that can improve prediction accuracy and achieve lower mean absolute error values compared with the results of CF or DF alone. To test our approach, we ran an offline evaluation using the 1M MovieLens data set.

Keywords: Recommender systems, Collaborative filtering, Demographic filtering, Hybrid system

1 Introduction

The amount of available information on the Internet has increased exponentially in the last decade and this has led to the problem of information overload. More specifically, the E-commerce industry is presenting a wider range of options, which makes it more difficult for users to shop and find products. Hence, to help customers find new items by means of suggestions, companies need to develop a

recommender system. Such systems help their sales to grow by providing relevant options that meet users requirements. For example, in regards to movie recommendation, the Netflix Prize raised the importance of the recommender system in attracting more users, and the competition to produce highly developed algorithms led to more accurate results in recommendations [2]. A recommendation system is an information filtering task that is used to predict the items a certain user will like (the prediction problem) or to recommend a set of top items that meet the users preferences [23]. Users have trouble handling large volumes of information, and problems with cognitive and data sparsity when attempting to find appropriate information at the right time [5]. Based on profile data, a recommender system can be categorized into four main stages: similarity computation, neighbourhood selection, prediction and recommendation. The profile can be modelled by content-based, collaborative or demographic filtering. If the user profile contains a set of attributes obtained from the item descriptions that the user has liked, this is content-based filtering (CBF). Demographic filtering (DF) is represented by a set of features in a users profile. Collaborative filtering (CF) can be described as when the profile contains a list of items that have been rated. The CF technique is widely used as a recommender system base method due to its capability and the efficiency of predicting similar neighbour users. However, extensive growth in the number of users and items may cause a sparsity issue in CF techniques when used on their own. Thus, we have made the following contributions:

1. We have developed a method that exploits both ratings and demographic information, by combining demographic attributes with user-item rating CF to solve the problem of data sparsity.
2. This method allows us to efficiently calculate similarities in a large dataset with no pre-calculating or pre-processing.
3. We have evaluated our method using a real dataset and we have shown that it is both practical and effective.

2 Related Work

Recommender systems were first researched in the mid-1990s, relying on the idea that users share similar items or opinions, thereby helping to make recommendations to others [20]. The researchers established a collaborative filtering technique based on ratings structure. Hence, the most common formulation is to calculate ratings for items that have not been seen by a particular user. Recommender systems can be defined as adaptable tools that help users search for, filter and classify information, and then recommend relevant items [22]. Recommendation systems use a number of different techniques. These methods can be implemented based on the domain requirement and are able to identify and predict items that meet the users interests. They also utilize different recommendation algorithms to make suggestions and recommendations.

2.1 Collaborative filtering

Collaborative filtering is considered to be the most popular technique for recommender systems. It has been widely implemented in different domains to make recommendations. It is a method of information filtering that seeks to predict the rating that a user would give to a particular item based on a similarity matrix. Collaborative filtering provided the foundation for the first recommender systems, which were used to help people make choices based on the opinions of other people [10]. It helps users to find relevant items and makes suggestions based on similar users tastes. It has been applied in a variety of areas, such as in regards to movies, books and research articles. In this approach the similarity calculation is based on the users peers. User-based CF: this method looks for similarity between users based on the same rating pattern [24]. It makes a recommendation based on the similarity between the target user and other users. The idea is that, for a given user, the preferences of similar users (neighbours) can serve as recommendations. A user-user approach was proposed as an appropriate method for recommending items based on expert opinions [3]. In addition, another example is provided by [30], where mobile activities were recommended to users based on their locations. Item-based CF: this method recommends items based on similarities between items shared with similar users [24]. In [14] an item to item collaborative filtering approach was designed that matched items rated or purchased by the user with other similar items.

2.2 Demographic filtering

It is possible to identify the type of person that likes a particular item by referencing their demographic details [18]. User attributes are incorporated into demographic recommender systems and this demographic data is used as a basis for arriving at suitable recommendations, sometimes relying on pre-generated demographic clusters [26]. This information is gathered either explicitly through user registrations or implicitly via navigation of the system they use [17]. Subsequently, demographically similar users are identified by means of the recommendation algorithm. Recommendations are based on how similar people (in terms of their demographics) rated a particular item [25]. In [26] a hybrid algorithm was presented that keeps the core ideas of two existing recommender systems and enhances them with relevant information extracted from demographic data. The authors in [18] presented an approach that considers user profiles as vectors constructed from demographic attributes such as age, gender or postcode to find relationships with other users and calculate similarities between users, in order to generate the final prediction. Demographic-based filters are similar to collaborative filters in the sense that both are able to identify similarities between users. In this case, demographic features are used to determine similarity rather than the users previous ratings of items [25]. Demographic attributes are added as meta-data to help the neighbourhood algorithm find similar users. The author in [16], presented the importance of these meta-data in producing significant results and providing better recommendations.

In [19], the authors stated that demographic information helps to address the cold-start problem. This is because this approach does not require a detailed history of user ratings before making recommendations, unlike the content-based and collaborative approaches [9]. studied the importance of demographic information (age, gender) in a research paper recommender system [4]. The authors showed that demographic information had a significant impact on recommendations. The combination of collaborative filtering and the demographic base can enrich user preferences and more accurately identify their interests.

2.3 Hybrid recommendation approaches

More recently, the hybrid recommendation approach has become a widely debated issue. A possible way to combine the recommendation methods was introduced in [8]. Authors in [28] also introduced a hybrid approach for solving the problem of finding the rating of unrated items in a user-item matrix through a weighted combination of user-based and item-based collaborative filtering. These methods addressed the two major challenges of recommender systems, the accuracy of recommendations and sparsity of data, by simultaneously incorporating the correlation of users and items.

Because of data sparsity, finding nearest neighbours is becoming more of a challenge, with the fast growth in users and items. In [1] a switching hybrid approach was proposed to solve the long tail problem in recommendations. A hybrid approach was applied that utilised clustering and genetic algorithms to reduce data sparsity in movie recommendations. The results showed that this approach improves recommendation accuracy [27]. In [7] a hybrid framework was proposed that utilized collaborative filtering relying on user/item metadata and demographic data. The framework benefits from the similarity between users via correlation in terms of demographic attributes. This improves prediction and is able to solve the cold start problem compared with the baseline. The author points out the importance of item metadata in overcoming the challenge in which user and item have little information. In [26] the discussion explored the usefulness of demographic data as an enhancing factor, by employing a hybrid algorithm to improve collaborative filtering in terms of both algorithms, user-based and item-based. In [11] a combination algorithm was proposed using demographic attributes based on a clustering approach in a weighted scheme. It solved the cold-start problem by assigning a new user to the nearest cluster based on demographic similarity. Our proposed method is beneficial for exploring the effect of demographic data and makes a comparison between the kNN and the Random Forest classifiers.

3 Proposed method

In this section, the proposed method is defined that combines CF and DF. The main idea of the method, which is not found in other works in the literature is to have a hybrid recommendation approach that can be easily used for the

evaluation of different classifiers in order to identify which classifier performs better when demographic data are integrated into the recommendation process.

The sparsity issue is a major challenge for recommender systems in terms of producing the right recommendations for the right users. This issue has been further expanded due to the growth of items available and of users with few ratings and little user information. This leads to difficulty in finding similarity between two users. In this section, we propose how a feature combination hybrid approach solves the sparsity problem and reduces the error rate through using two classifiers. It combines matching user demographic attributes with the user rating CF method as shown in Figure 1.

In order to evaluate the proposed method we conducted an experiment on a real dataset that is publicly available from MovieLens [12]. In this paper we used the 1M dataset which contains 1,000,209 ratings that were assigned by 6,040 users on around 3,900 movies. We utilized demographic information that includes age, gender and occupation. We combined the demographic information for each user with user-item ratings. Hence, each user was defined as a vector with those features.

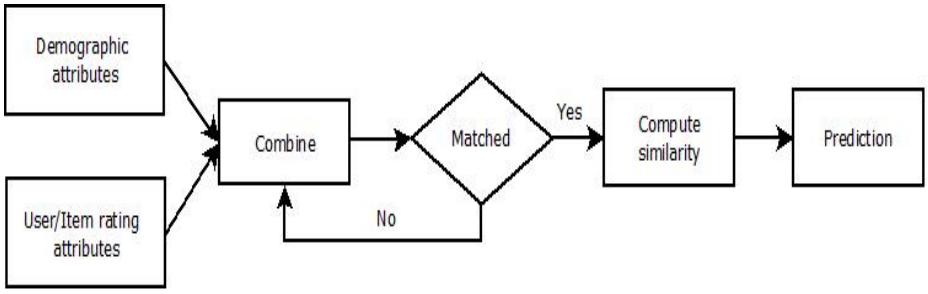


Fig. 1: The Architecture of the proposed method.

The attributes used in this filtering system are age, gender and occupation. Those attributes are defined as categorical, and represent each user in a group. They can help in finding similar users, in order to improve rating prediction accuracy. The profile vector is represented at the attribute-level to compute the similarity.

Then, we calculated the similarity between the active user and the nearest one. Next is the final step of calculating the predicted rating. We ran this experiment using Orange 3.7.0, which is a data mining and machine learning tool. We conducted a cross-validation with number of folds = 10. In summary, the steps of the proposed method are:

1. CF is combined with demographic attributes such as age, gender and occupation to find more similar users. The combination was made through matching the user ID from user-item ratings data with user demographics data as detailed in algorithm 1 below. Where row[0] and line[0] represent

user-id. And, row[0,1,..,N] represent the attributes in CF and line[1,2,...,N] represent the attributes in DF.
2. After matching each user with the demographic attributes, the similarity is computed using two classifier kNN and Random Forest.
3. The final step is, predicted rating is calculated and compared with the actual rating to calculate the differences.

Algorithm 1 Combined algorithm

```
Input: user-item rating attributes file (f1). demographic attributes (f2).
Output: user demographic attributes with item rating (f3).
for <row in f2> do
    for <line in f1> do
        if row [0] == line [0] then
            f3 = row [0,1,..,N] + line [1,2,...,N]
        end if
    end for
end for
```

4 Experimental Evaluation

Evaluation metrics play an important role in measuring the quality and performance of a recommender system. Since 1994 [21], the accuracy of the recommender system has been evaluated in the literature in different ways. Furthermore, as there is no standard for evaluation, it is hard to compare the results with other published articles. However, there are main evaluation metrics that are widely applied to benchmark the results and compare them with the proposed algorithms. Most of the empirical studies examining recommender systems have focused on appraising the accuracy of these systems [13]. This insight is useful for evaluating the quality of the system and its ability to forecast the rating for a particular item.

4.1 Predictive accuracy metrics

This measures similarity between true user ratings and the predicted ratings. This research applies accuracy metrics to measure the performance of the proposed methods. Both the mean absolute error (MAE) in Equation 1 and the root mean squared error (RMSE) in Equation 2 is used to evaluate the prediction accuracy of the different recommendation techniques, where p_i is the predicted rating and r_i is the actual rating. Prediction accuracy is enhanced when MAE and RMSE are lower. Here we detail those similarity measures

Mean Absolute Error (MAE) takes the mean of the absolute difference between the predicted rating and actual rating for all the ratings as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \tag{1}$$

Root Mean Squared Error(RMSE) represents the sample standard deviation of the differences between predicted values and the actual values:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2} \tag{2}$$

4.2 Classification algorithms

In order to find out which classifier is the most appropriate one to use for this dataset and to make a good prediction for the movie domain, we ran an experiment into those two that are widely applied in movie recommendation for evaluating the results. Next, we describe in detail each classifier.

k-Nearest Neighbour (kNN) classifier finds its k nearest neighbours. The given user is assigned to a similar users that shares the most common features of its k nearest neighbour users. Certain factors need to be considered, such as the similarity measurement, which calculates the distance between two vectors p_i and r_i for i = 1,2,...,k representing the neighbourhood, which needs to be positive number. For that purpose, we use Euclidean distance as follows:

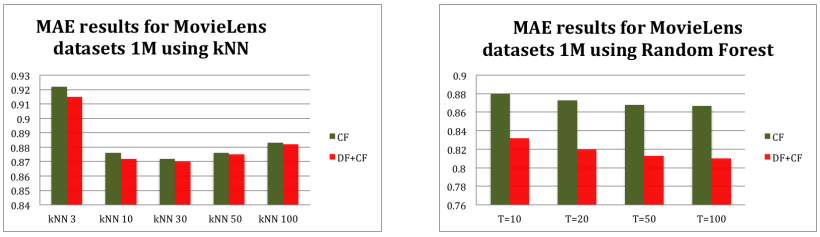
$$\mathrm{d(p,q)} = \sqrt{\sum_{i=1}^{k}(p_i - r_i)^2} \tag{3}$$

Random Forest is an ensemble learning classifier that builds a set of decision trees. Each tree is developed from a bootstrap sample from training data. It is more robust with respect to noise [6]. This method has been successfully approved as an accurate machine learning classifier [15]. We set the number of trees to be 10, 20, 50 and 100, which is the most likely change between this range [29].

4.3 Results

As we can see in Figure 2, the MAE accuracy metrics were made through applying different kNN with k = 3, 10, 30, 50 and 100. We then performed the experiment with the Random Forest classifier using a different set of trees. It is clear that performance is improved when we combined the demographic attributes (CF+DF). However, it is noticeable that the improvement in Random

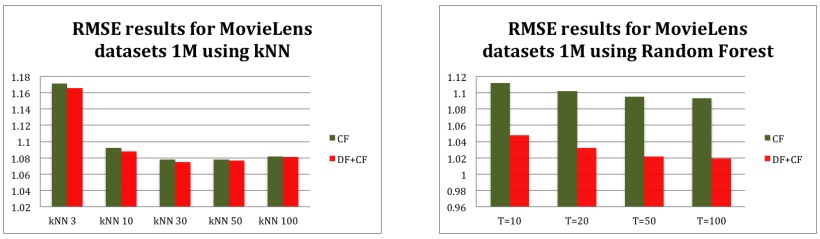
Forest is much higher than in kNN. For example, in Figure 2b, when $T = 10$ the improvement is 5%. Whereas in Figure 2a, when $k = 3$ the improvement is only 1%.



(a) kNN (b) Random Forest

Fig. 2: MAE results for the MovieLens dataset using kNN and Random Forest

Figure 3 shows the results using RMSE. There is significant improvement in the Random Forest compared to kNN in all sets. For instance, in Figure 3b when $T = 10$ the collaborative filtering performed 1.11 whereas the combined method it was 1.05. By contrast, in figure 3a when $k = 3$ the enhancement is only 0.01.



(a) kNN (b) Random Forest

Fig. 3: RMSE results for the MovieLens dataset using kNN and Random Forest

In general, the Random Forest significantly outperforms the kNN. Therefore, this experiment proves the quality of the Random Forest compared with the kNN classifier using this large dataset. We also ran all the data to make a reliable evaluation and produce an accurate result, and compared it to the benchmark. Our hybrid method outperformed the baseline CF method.

5 Conclusion

In this paper, we propose a novel hybrid method for recommender systems based on simultaneous combination of user-based collaborative filtering and demographic attributes. The results suggest that demographic filtering can effectively

improve the overall recommendation. Moreover, the proposed method addresses two common challenges of recommendation systems, namely sparsity of data and improved accuracy of recommender systems, by combining the hidden relations between users and comparing two different classifiers with a large dataset. The proposed method is a comparison between the Random Forest and kNN classifiers. In future work, we may consider specific users who rate only a few items, or possibly other attributes relating to item representation. In addition, other different classifiers could be used to add a comprehensive comparison.

References

1. Alshammari, G., Jorro-Aragoneses, J.L., Kapetanakis, S., Petridis, M., Recio-García, J.A., Díaz-Agudo, B.: A hybrid cbr approach for the long tail problem in recommender systems. In: International Conference on Case-Based Reasoning. pp. 35–45. Springer (2017)
2. Amatriain, X.: Beyond data: from user information to business value through personalized recommendations and consumer science. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 2201–2208. ACM (2013)
3. Amatriain, X., Lathia, N., Pujol, J.M., Kwak, H., Oliver, N.: The wisdom of the few: a collaborative filtering approach based on expert opinions from the web. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 532–539. ACM (2009)
4. Beel, J., Langer, S., Nürnberger, A., Genzmehr, M.: The Impact of Demographics (Age and Gender) and Other User-Characteristics on Evaluating Recommender Systems. In: Research and Advanced Technology for Digital Libraries, pp. 396–400. Springer Science $$ Business Media (2013), http://dx.doi.org/10.1007/978-3-642-40501-3{_}45
5. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Knowledge-Based Systems Recommender systems survey 46, 109–132 (2013)
6. Breiman, L.: RANDOM FORESTS pp. 1–33 (2001)
7. Bremer, S., Schelten, A., Lohmann, E., Kleinsteuber, M.: A Framework for Training Hybrid Recommender Systems pp. 30–37 (2017)
8. Burke, R.: Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction 12(4), 331–370 (2002)
9. Burke, R.: The adaptive web. chapter hybrid web recommender systems (2007)
10. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of the ACM 35(12), 61–70 (1992)
11. Gupta, J.: Performance Analysis of Recommendation System Based On Collaborative Filtering and Demographics (2015)
12. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5(4), 19 (2016)
13. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 230–237. ACM (1999)
14. Linden, G., Smith, B., York, J.: Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet computing 7(1), 76–80 (2003)

15. Louppe, G.: Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502 (2014)
16. Mittal, P.: Metadata Based Recommender Systems pp. 2659–2664 (2014)
17. Moldovan, A.N., Muntean, C.H.: Personalisation of the multimedia content delivered to mobile device users. In: Broadband Multimedia Systems and Broadcasting, 2009. BMSB'09. IEEE International Symposium on. pp. 1–6. IEEE (2009)
18. Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: The adaptive web, pp. 325–341. Springer (2007)
19. Redpath, J.L.: Improving the performance of recommender algorithms. Ph.D. thesis, University of Ulster (2010)
20. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens : An Open Architecture for Collaborative Filtering of Netnews. Proceedings of the 1994 ACM conference on Computer supported cooperative work pp. 175–186 (1994)
21. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM conference on Computer supported cooperative work. pp. 175–186. ACM (1994)
22. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM 40(3), 56–58 (1997)
23. Shah, L., Gaudani, H., Balani, P.: Survey on Recommendation System 137(7), 43–49 (2016)
24. Spiegel, S.: A hybrid approach to recommender systems based on matrix factorization. Department for Agent Technologies and Telecommunications, Technical University Berlin (2009)
25. Tintarev, N.: Explaining recommendations. User Modeling 2007 pp. 470–474 (2009)
26. Vozalis, M., Margaritis, K.G.: Collaborative filtering enhanced by demographic correlation. In: AIAI symposium on professional practice in AI, of the 18th world computer congress (2004)
27. Wang, Z., Yu, X., Feng, N., Wang, Z.: Journal of Visual Languages and Computing An improved collaborative movie recommendation system using computational intelligence $ 25, 667–675 (2014)
28. Wei, S., Zheng, X., Chen, D., Chen, C.: Electronic Commerce Research and Applications A hybrid approach for movie recommendation via tags and ratings q 18, 83–94 (2016)
29. Zhang, H., Min, F., Wang, S.: A random forest approach to model-based recommendation. JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE 11(15), 5341–5348 (2014)
30. Zheng, V.W., Cao, B., Zheng, Y., Xie, X., Yang, Q.: Collaborative filtering meets mobile recommendation: A user-centered approach. In: AAAI. vol. 10, pp. 236–241 (2010)