Electronic Commerce Research and Applications 29 (2018) 12–18

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/ecra

Assessing new correlation-based collaborative filtering approaches for binary market basket data

Wook-Yeon Hwang

College of Global Business, Dong-A University, 49230 SeoGu GudukRo 225, Busa, South Korea

ARTICLE INFO

Article history:
Received 10 February 2018
Received in revised form 4 March 2018
Accepted 4 March 2018
Available online 10 March 2018

Keywords:
Binary market basket data
Collaborative filtering
Pearson correlation
Top-*N* accuracy

ABSTRACT

Binary market basket data are common in marketing settings. Pearson correlation-based approaches have applied to this kind of data in the past. This research assessed the principles of this approach research identifies some related problems. By resolving the problems, I develop new Pearson correlation-based approaches that use separated terms and separated terms with proportions. The experimental results show that the approaches perform better than the existing ones for market basket data in terms of top-*N* accuracy.

1. Introduction

Recommender systems are widely used for industry marketers when they make contact with a customer because the systems effectively recommend products which customers like. A variety of collaborative filtering (CF) approaches have been utilized for recommender systems (Park et al., 2012, Su and Khoshgoftaar, 2009). The approaches are classified into user-based CF with user-item similarities and model-based CF (Breese et al., 1998). Among the user-based CFs that use user-item similarities, those that leverage Pearson correlation (hereafter without referring to "Pearson" anymore) have been applied to scoring data sets because they are easy to use.

Breese et al. (1998) first proposed user-based CF to leverage correlation. However, user-based CF does provide satisfactory accuracy because a majority of the voting scores are missing (Ahn, 2008). Instead of using original data sets, Mild and Reutterer (2001) proposed to apply user-based CF to tbinary market basket data, which are represented by ones (purchased) and zeros (non-purchased) (Yang and Lai, 2006). However, the performance also was not satisfactory. Thus, they developed the correlation-based approach instead (Mild and Reutterer, 2003). I focus on the correlation-based approach for the binary market basket data to reveal more knowledge about it properties and performance.

For binary market basket data, in order to improve prediction performance when the number of training users is enough, supervised learning approaches have been applied (Lee et al., 2005, Lee and Olafsson, 2009, Hwang and Jun 2014, Hwang, 2018) based on binary classification and regression modeling. Lee et al. (2005) and Lee and Olafsson (2009) considered logistic regression with principal components for market basket data (Hastie et al., 2001). Hwang and Jun (2014) utilized the random forest and elastic net approaches to tackle the problem (Breiman, 2001, Zou and Hastie, 2005). Hwang (2018) proposed two new variable selection approaches: a correlation-based variable selection approach (VS1) and a forward random forest regression-based variable selection approach (VS2).

In binary market basket data, I can identify respondents and non-respondents, where zeros represent respondents and ones represent the non-respondents. Therefore, the CF approaches for the binary market basket data will be useful when marketers want to contact respondents at some cost, as in targeted marketing. Likewise, Lee et al. (2010) considered response modeling for classifying customers who are likely to purchase a promoted product and customers who are not. In general, the correlation-based approach is more effective than supervised learning when the number of training users is not enough (Lee et al., 2005, Hwang and Jun 2014, Hwang, 2018). This is the cold-start problem, and is why I wish to improve the correlation-based approach in this research.

Though the correlation-based approach is simple and easy to use, its principles and problems for handling binary market basket data have not been investigated carefully. In this research, I analyze the principles of the correlation-based approach and identify its problems. In particular, this approach not only ignores information on non-purchased items, but also sticks to the unsystematic voting score function. Moreover, the correlation coefficient cannot

E-mail address: wyhwang@dau.ac.kr

https://doi.org/10.1016/j.elerap.2018.03.002
1567-4223/

represent the contributions of the binary combinations very well. By resolving these problems, I hope to improve the correlation-based approach. As a result, I propose new approaches to tackle binary market basket data-based recommendation.

2. Pearson correlation-based approaches

The correlation-based approach depends on user-item similarities, where the user-based CF is considered. A binary variable, v_{ij} is defined, where $\boldsymbol{V} = (v_{ij})$ represents a binary user-item matrix (training data), $i = 1,2, \ldots, n$; $j = 1,2, \ldots, m$), where ones represent purchased items and zeros non-purchased items. Based on Mild and Reutterer's approach (2003), Lee et al. (2005) denoted the predicted voting score for an active user a for the item j, P_{aj}, by:

$$P_{aj} = k_a \sum_{i=1}^{n} w(a,i)\, v_{ij}, \tag{1}$$

where $\bar{v}_i = \frac{1}{m}\sum_j v_{ij}$, $\bar{v}_a = \frac{1}{m}\sum_j v_{aj}$, $w(a,i) = \frac{\sum_j (v_{aj}-\bar{v}_a)(v_{ij}-\bar{v}_i)}{\sqrt{\sum_j (v_{aj}-\bar{v}_a)^2 \sum_j (v_{ij}-\bar{v}_i)^2}}$, and $k_a = \frac{1}{\sum_{i=1}^{n} |w(a,i)|}$.

Instead of a binary user-item matrix, I consider a binary item-user matrix. In this context, item-user similarities can be considered for the item-based CF (Lee and Olafsson 2009, Hwang and Jun 2014, Hwang, 2018). The correlation-based approach is better than the user-based CF leveraging of the correlation for the binary user-item matrix (Mild and Reutterer, 2003).

To improve the original correlation-based approach, Hwang (2018) developed a new variable selection approach for the correlation-based approach, which is called VS1 modeling (Variable Selection 1). The modified Eq. (1) is as follows:

$$P_{aj} = k_a \sum_{i \in S(a)} w(a,i)\, v_{ij}, \tag{2}$$

where $S(a) = \{i \,||w(a,i)|' \text{srank} \leqslant c \text{and} a \neq i\}$, for some integer c between 1 and n. In order to decide $S(a)$, $|w(a,i)|$ is ranked in descending order, some integer c can represent only those users who are highly correlated with the active user for $S(a)$. One disadvantage of VS1 modeling is that it may not find a reasonable value of c to improve prediction performance for the user-based CF, which means that the correlation-based approach cannot be improved upon.

Table 1
Possible contributions of $\boldsymbol{w}(\boldsymbol{a},\boldsymbol{i})$ in Eq. (1).

v_{ij}	The sign of w(a, i)	Contributions of w(a, i)
$v_{ij} = 1$	Positive sign ($w(a,i) > 0$)	Adding $w(a,i)$ to $\sum_{i=1}^{n} w(a,i)\, v_{ij}$
	Negative sign ($w(a,i) < 0$)	Subtracting $w(a,i)$ from $\sum_{i=1}^{n} w(a,i)\, v_{ij}$
$v_{ij} = 0$	Positive sign ($w(a,i) > 0$)	Subtracting $w(a,i)$ from $\sum_{i=1}^{n} w(a,i)\, v_{ij}$
	Negative sign ($w(a,i) < 0$)	Adding $w(a,i)$ to $\sum_{i=1}^{n} w(a,i)\, v_{ij}$

3. New Pearson correlation-based approaches

3.1. Problems with the approach

3.1.1. Structural issues

In Eq. (1), for $v_{ij} = 1$ and $w(a,i) > 0$, $w(a,i)$ should be added to $\sum_{i=1}^{n} w(a,i)\, v_{ij}$. This is because a purchase of the item is expected by the active user, when the correlation is positive and the item is purchased by the training user. In contrast, for $v_{ij} = 1$ and $w(a,i) < 0$, $w(a,i)$ should be subtracted from $\sum_{i=1}^{n} w(a,i)\, v_{ij}$. This is because a purchase of the item is not expected by the active user, when the correlation is negative and the item is purchased by the training user. On the other hand, the predicted voting score for $v_{ij} = 0$ is not counted at all. This is due to $\sum_{i=1}^{n} w(a,i)\, v_{ij}$ for $v_{ij} = 0$ always.

The predicted voting score for $v_{ij} = 0$ should not be ignored, however. This is because it can contribute to the prediction. Table 1 elaborates on the possible contributions of $w(a,i)$ for the predicted voting score given the value of v_{ij} and the sign of $w(a,i)$. As discussed above, if the sign of $w(a,i)$ for $v_{ij} = 1$ is positive, then $w(a,i)$ should be added to $\sum_{i=1}^{n} w(a,i)\, v_{ij}$, whereas $w(a,i)$ should be subtracted from $\sum_{i=1}^{n} w(a,i)\, v_{ij}$ if the sign of $w(a,i)$ for $v_{ij} = 1$ is negative.

Similarly, if the sign of $w(a,i)$ for $v_{ij} = 0$ is positive, $w(a,i)$ should be subtracted from $\sum_{i=1}^{n} w(a,i)\, v_{ij}$ because a purchase of the item is not expected by the active user, when the correlation is positive and the item is not purchased by the training user. On the other hand, if the sign of $w(a,i)$ for $v_{ij} = 0$ is negative, $w(a,i)$ should be added to $\sum_{i=1}^{n} w(a,i)\, v_{ij}$ because a purchase of the item is expected by the active user, when the correlation is negative and the item is not purchased by the training user. Because the correlation-based approach ignores the contributions of the predicted voting scores for $v_{ij} = 0$, it should be modified by considering the contributions it makes.

Moreover, in Eq. (1), the correlation-based approach simultaneously considers both the positive and negative correlations for $v_{ij} = 1$. However, it is more efficient to choose only one when only positive correlations or only negative correlations are the main contributors. Therefore, I separately consider positive correlation from negative correlation for $v_{ij} = 1$. For instance, if the contribution of negative correlation for $v_{ij} = 1$ is negligible, I should choose the positive correlation only. Instead of a binary user-item matrix, I consider a binary item-user matrix (Lee and Olafsson 2009, Hwang and Jun 2014). Similarly, as shown in Table 1, it is possible to conduct some reasoning for possible contributions in the predicted voting score for the binary item-user matrix.

3.1.2. Limitations

I consider a contingency table for a training user and an active user in the binary market basket data. The frequencies for the four combinations, (0,0), (0,1), (1,0) and (1,1) are counted as shown in Table 2. The frequencies can represent the contributions of the four combinations to the correlation coefficient. For instance, an

Table 2
Contingency table for a training user and an active user.

		Training user		
		0	1	Sum
Active user	0	"0 0"= Freq.for (0, 0)	"0 1"= Freq.for (0, 1)	"0 0" + "0 1"
	1	"1 0" = Freq.for (1, 0)	"1 1"= Freq.for (1, 1)	"1 0"+"1 1"
	Sum	"0 0" + "1 0"	"0 1" + "1 1"	"0 0" + "0 1"+"1 0" + "1 1"