An Optimal Scaling Approach to Collaborative Filtering using Categorical Principal Component Analysis and Neighborhood Formation

Angelos I. Markos[1], Manolis G. Vozalis[2], and Konstantinos G. Margaritis[2]

[1] Department of Primary Education, Democritus University of Thrace
amarkos@eled.duth.gr
[2] Department of Applied Informatics, University of Macedonia
{mans,kmarg}@uom.gr

Abstract. Collaborative Filtering (CF) is a popular technique employed by Recommender Systems, a term used to describe intelligent methods that generate personalized recommendations. The most common and accurate approaches to CF are based on latent factor models. Latent factor models can tackle two fundamental problems of CF, data sparsity and scalability and have received considerable attention in recent literature. In this work, we present an optimal scaling approach to address both of these problems using Categorical Principal Component Analysis for the low-rank approximation of the user-item ratings matrix, followed by a neighborhood formation step. The optimal scaling approach has the advantage that it can be easily extended to the case when there are missing data and restrictions for ordinal and numerical variables can be easily imposed. We considered different measurement levels for the user ratings on items, starting with a multiple nominal and consecutively applying nominal, ordinal and numeric levels. Experiments were executed on the MovieLens dataset, aiming to evaluate the aforementioned options in terms of accuracy. Results indicated that a combined approach (multiple nominal measurement level, "passive" missing data strategy) clearly outperformed the other tested options.

Keywords: Collaborative filtering, Recommender systems, Low-rank approximation, Categorical Principal Component Analysis, Optimal scaling

1 Introduction

Recommender Systems (RSs) are intelligent applications that generate personalized recommendations. RSs often rely on Collaborating Filtering (CF), which is based on the premise that users who have agreed in the past, tend to agree in the future. Among the most common and successful approaches to CF are latent factor models, which directly profile both users and products. The goal of latent factor models is to uncover latent features that explain user preferences by transforming both items and users to the same latent factor space, thus making them directly comparable. Latent factors represent either obvious characteristics of the items or completely uninterpretable dimensions. Each factor

measures the degree of preference for each user towards items that score high on the corresponding factor.

Several latent factor models have been successfully applied to CF. Sarwar et al. [1] apply Singular Value Decomposition to reduce the dimensionality of sparse rating matrices and improve the scalability of CF systems. Goldberg et al. [2] use Principal Component Analysis (PCA) as a preprocessing step for low-rank approximation and then utilize clustering to predict user preferences. Hoffman [3] suggests a collaborative prediction method based on a probabilistic latent variable model. Salakhutdinov and Mnih [4] propose a momentum based, probabilistic matrix factorization algorithm with batch learning which scales linearly with the number of observations. Bell and Koren [5] present a neighborhood based approach, which uses alternating least squares and removes the global effect from the data. Tacacs et al. [6] discuss and compare modifications of already published matrix factorization methods. In addition to a straightforward optimization approach for the general low-rank approximation problem, many authors propose extensions of well established algorithms to cope with the missing data problem [7, 8].

In this work, we present an optimal scaling approach to address two fundamentals problems of CF, data sparsity and scalability. Optimal scaling is a general approach to treat multivariate data through the optimal transformation of qualitative scales to quantitative values [9, 10]. In other words, both nominal and ordinal variables (or features) can be optimally transformed to variables with numeric properties. These optimal transformations of the original variables may be used to overcome the linear assumption underlying many classic dimensionality reduction methods. In this light, PCA has been reviewed and extended to Categorical or Nonlinear Principal Component Analysis (CatPCA) with many potential applications [11].

Our approach uses CatPCA for the low-rank approximation of the user-item ratings matrix, followed by a neighborhood formation step. The combination of matrix factorization and neighborhood formation can lead to accurate predictions [5, 6]. CatPCA is based on the Alternating Least Squares (ALS) algorithm utilizing an optimal least squares scaling process where original data are transformed so that their overall variance is maximized. The problem is formulated by means of a loss function and it is solved by the ALS. Eventually, this leads to optimally scaled scores on each factor [12]. Unlike other matrix factorization methods, CatPCA does not assume multivariate normality and linear relationships between variables and provides a flexible framework for parametrization. In particular, we consider different measurement levels for the user ratings on items, starting with a multiple nominal and consecutively applying nominal, ordinal and numeric levels. A series of experiments was executed on the MovieLens dataset in order to evaluate the aforementioned options in terms of accuracy. A combined approach (multiple nominal level, "passive" missing data strategy) clearly outperformed the other tested options.

The remainder of this paper is organized as follows. The next section, Section 2, gives a brief presentation of CatPCA with optimal scaling. In Section 3,

the proposed approach is thoroughly described, starting with the low-rank approximation of the ratings matrix, which is achieved through CatPCA, and continuing with a user neighborhood formation step. The efficiency of four distinct variations of this approach is demonstrated in Section 4 through a set of experiments on a publicly available dataset. Finally, we give conclusions and provide a discussion over the results in Section 5.

2 Categorical Principal Component Analysis with Optimal Scaling

CatPCA with optimal scaling or optimal scoring is a general approach to treat multivariate data through the optimal transformation of qualitative scales to quantitative values [9, 10]. In the optimal transformation process, an appropriate quantification level has to be chosen for each of the variables. The most restricted transformation level is called *numeric*; it applies a linear transformation to the original integer scale values, so that the resulting variables are standardized. The order of variable categories and the equal distances between category numbers of the original variable are preserved in the optimally scaled variable. When all variables are at a numeric analysis level, the analysis is analogous to standard PCA. The optimal least squares transformations of *ordinal* data can be handled by means of monotonic transformations, which maintain the order in the original data. In the case of *nominal* data, categories are not ordered and the only information that is preserved is the grouping of objects in categories. Finally, when all variables are at a *multiple nominal* level, the only information that is preserved is the grouping of objects in categories. The quantification is called *multiple* because there is a separate quantification for each dimension. With all variables at a multiple nominal level, the analysis is equivalent to multiple correspondence analysis or homogeneity analysis [9, 13].

3 Algorithm Description

In this Section we describe the proposed optimal least squares approach for CF, which starts with the low-rank approximation of the user-item ratings matrix and is followed by a user neighborhood formation step.

3.1 Low-Rank approximation through CatPCA

We start with the following basic definitions. For $i = 1, \ldots, n$ users, ratings on $j = 1, \ldots, m$ items are collected in the $n \times m$ data matrix $\mathbf{R}$. Each of the corresponding items takes on k different rating values (levels or categories) from a given range, i.e. $(1, 2, 3, 4, 5)$. They are coded using m binary indicator or dummy matrices $\mathbf{G}_j$ of size $n \times k$, with entries $\mathbf{G}_j(i, t) = 1$ if user i has given item j a rating of t, and $\mathbf{G}_j(i, t) = 0$ for any of the remaining rating values. The whole set of indicator matrices can be collected in a block matrix $\mathbf{G} = [\mathbf{G}_1 \ldots \mathbf{G}_m]$.

Missing item ratings are coded as complete zero row sums: if user i has not rated item j, then the row sum of $\mathbf{G}_j$ is 0. Otherwise the row sum becomes 1. Let $\mathbf{M}_j$ denote the $n \times n$ binary diagonal matrix with entries $\mathbf{M}_j(i,i) = 1$ if user i has rated item j and 0 otherwise. Based on $\mathbf{M}_j$ we define $\mathbf{M}_\star$ as the sum of the $\mathbf{M}_j$'s and $\mathbf{M}_\bullet$ as their average. This missing data strategy is known in the literature as *missing data passive*, because it leaves the indicator matrix $\mathbf{G}_j$ incomplete [12]. The main advantage of this option is that it ignores the missing ratings without making strong assumptions regarding the pattern of missing data.

Let $\mathbf{X}$ be the unknown $n \times p$ matrix containing the user scores in the latent factor space R^p and $\mathbf{Y}_j$ be the unknown $k \times p$ matrix containing the rating scores (quantifications) of item j in the same p-dimensional space. Both matrices have to be determined during optimization. Based on these definitions, the following loss function can be established:

$$\sigma(\mathbf{X}; \mathbf{Y}_1 \dots \mathbf{Y}_m) = \frac{1}{m} \sum_{j=1}^{m} tr(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)'\mathbf{M}_j(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j) \tag{1}$$

which is optimized under the normalization conditions $u'\mathbf{M}_\bullet\mathbf{X} = 0$, with u denoting an n vector with ones and $\mathbf{X}'\mathbf{M}_\bullet\mathbf{X} = I$ such that the trivial solution of complete 0-scores is avoided. From an analytical point of view the loss function represents the sum of squares of $(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)$ and a simultaneous minimization over $\mathbf{X}$ and $\mathbf{Y}_j$. Minimizing equation 1 leads to a low dimensional space where users are positioned as close as possible to the ratings they have given to specific items and where an item's rating is the centroid of the users who have given this rating to the item.

The minimization problem can be solved by means of an iterative ALS procedure as follows:
At iteration 0 we begin with a random starting solution $\mathbf{X}^0$ for the user scores. Each iteration s consists of three steps:

1. Update item rating quantifications: $\mathbf{Y}_j^{(s)} = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{X}^{(s)}$ for $j = 1, \dots m$, where $\mathbf{D}_j = \mathbf{G}_j'\mathbf{G}_j$ the $k \times k$ diagonal matrix with the (marginal) frequencies of item j in its main diagonal.
2. Update user scores: $\widetilde{\mathbf{X}}^{(s)} = \mathbf{M}_*^{-1} \sum_{j=1}^{m} \mathbf{G}_j\mathbf{Y}_j^{(s)}$
3. Normalization: $\mathbf{X}^{(s+1)} = \mathbf{M}_*^{-1/2}\mathbf{orth()}(\mathbf{M}_*^{-1/2}\widetilde{\mathbf{X}}^{(s)})$
   Here **orth()** denotes the modified Gram-Schmidt process which is used to compute an orthonormal basis for the column space of a matrix.

This iterative process is continued until the improvement in subsequent loss values is below a convergence criterion. Note that matrix multiplications using indicator matrices can be implemented efficiently as cumulating the sums of rows over $\mathbf{X}$ and $\mathbf{Y}$. The ALS algorithm only computes the first p dimensions of the solution, which leads to an increase in computational efficiency. Moreover, by capitalizing on sparseness of $\mathbf{G}$, the algorithm is able to handle large datasets.

The approach described above is known as homogeneity analysis (homals) with all item ratings at a *multiple nominal* level of measurement and it has the advantage that it can be easily extended to the case when restrictions can be easily imposed. Nominal, numerical and ordinal item ratings can be incorporated with a rank-1 restriction of the form

$$\mathbf{Y}_j = \mathbf{z}_j \mathbf{a}_j' \tag{2}$$

where $\mathbf{z}_j$ is a column vector of length k with item rating quantifications and $\mathbf{a}_j$ a vector of length p with weights. Using this restriction the ratings of a specific item are forced to be on a line in the p dimensional space. However, this restriction is not sufficient to preserve the order for ordinal ratings or even the relative distance for numerical ratings. Therefore, $\mathbf{z}_j$ should be transformed in every ALS iteration to satisfy these restrictions. In the case of ordinal ratings this transformation is achieved by means of a weighted monotone regression of $\mathbf{z}_j$ on $\mathbf{r}_j$ and in the case of numerical ratings $\mathbf{z}_j$ is replaced by the original item rating $\mathbf{r}_j$. For a more detailed description we refer to [12, 9].

3.2 User Neighborhood Formation

At this point, we use the $n \times p$ matrix $\mathbf{X}$, which is obtained from the previous step, in order to form each user's neighborhood. Note that $\mathbf{X}$ contains the user scores in the p-dimensionality space. To find the proximity between two users, a and i, we utilize the Pearson product-moment correlation coefficient, as a similarity measure between each user and his closest neighbors. This similarity measure has been evaluated in a number of studies and has been found to be reliable and accurate. The coefficient is calculated as follows:

$$cor_{ai} = \frac{\sum_{j=1}^{l} r_{aj} r_{ij}}{\sqrt{\sum_{j=1}^{l} r_{aj} \sum_{j=1}^{h} r_{ij}}} \tag{3}$$

where r_{ij} denotes the rating of user i on item j. The summations over j are calculated over the l items for which both users a and i have expressed their opinions.

After a user neighborhood of size h has been formed for user a, we proceed with prediction generation. A prediction rating p_{aj} for user a on item j is computed using the following equation:

$$p_{aj} = \frac{\sum_{i=1}^{h} r_{ij} * cor_{ai}}{\sum_{i=1}^{h} |cor_{ai}|} \tag{4}$$

4 Experiments

In this section we provide a brief description of the experiments we executed in order to evaluate and compare the proposed methods, and then we present and comment on their results.

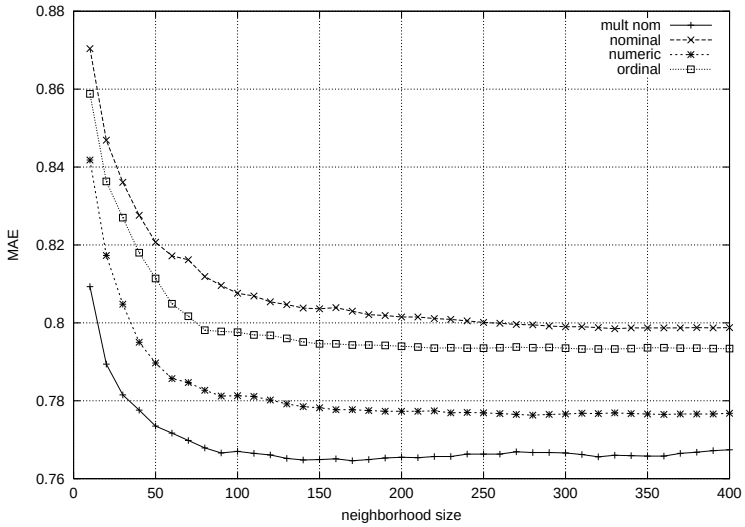

Fig. 1. Comparison of the three CatPCA algorithms for varying neighborhood sizes

For this purpose, we utilized the MovieLens dataset. It consists of 100,000 ratings which were assigned by 943 users on 1682 movies and is available at http://www.grouplens.org/. Ratings follow the 1(bad)-5(excellent) numerical scale. The sparsity of the dataset is high, at a value of 93.7%. Starting from the original dataset, five distinct splits of training (80%) and test (20%) data were utilized. The proposed algorithms were applied on each split and the average result across all 5 trials was computed.

Mean Absolute Error (MAE) was the metric we employed to evaluate the accuracy of the methods. MAE measures the deviation of predictions generated by the RSs from the true rating values, as they were specified by the user.

For our first experiment, we kept a fixed number of retained dimensions and tried to evaluate the impact of a varying neighborhood size on prediction accuracy. Figure 1 depicts the MAE values for neighborhood sizes ranging between 1 and 400 users. Based on that figure, it is clear that the multiple nominal and numeric approaches performed better than the nominal and ordinal ones. In particular, *multiple nominal* has generated the most accurate prediction among those tested (MAE=0.7646). This value was achieved for a neighborhood of 170 users, the smallest in size when compared to the number of neighbors required by the rest of the approaches to reach their optimal predictions. Specifically, the lowest MAE values for the numeric, nominal and ordinal approaches were 0.7763, 0.7985, and 0.7993, observed for neighborhoods including 280, 330 and 330 users, respectively.

Once the effect of varying neighborhood sizes was assessed, a second experiment was executed, aiming to evaluate the impact of dimensionality reduction on prediction accuracy. Figure 2 depicts the MAE results for values of retained

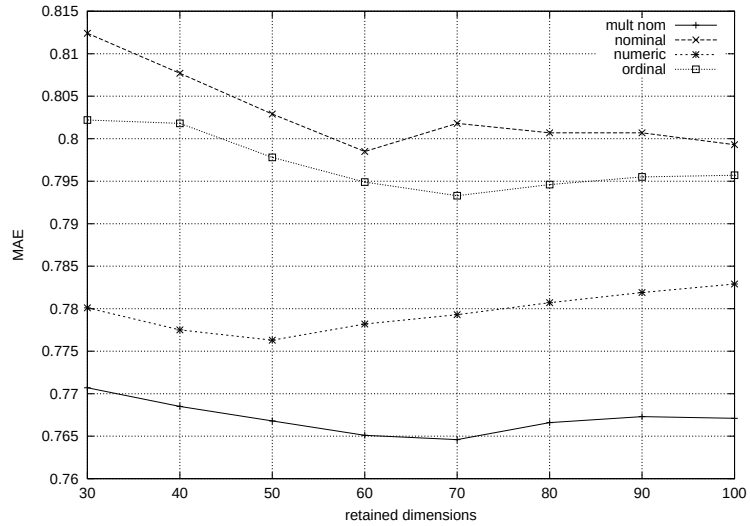

Fig. 2. Comparison of the three CatPCA algorithms for different values of retained dimensions

dimensions d, between 30 and 100. Based on that Figure, the advantage of *multiple nominal* approach over the rest was verified. Furthermore, both multiple nominal and ordinal levels reached their best MAE values for d=70, whereas d was 50 and 60 for numeric and nominal levels, respectively.

5 Conclusions

In this paper, we describe an optimal scaling approach for CF using CatPCA, which combines the low-rank approximation of the user-item ratings matrix with user neighborhood formation. We considered four different measurement levels for the user ratings in the Movielens dataset. A combined approach, where the user ratings on each item were handled as multiple nominal, gave the most accurate predictions according to MAE.

CatPCA with optimal scaling can offer a versatile set of options suited to CF problems. The main advantage of the proposed approach stems from the fact that CatPCA is able to account for more of the variance in the data compared to linear PCA, when the variables are (or may be) nonlinearly related to each other. Another advantage of this approach is that the "passive" strategy for handling missing data, ignores missing ratings without making strong assumptions for their pattern as is the case for various imputation methods. Additionally, predictions for new users of the system can be obtained by the model without requiring a complete retraining of the whole dataset. This is achieved after optimization by means of a cone restricted SVD [13], without influencing the existing scores computed by ALS.

In order to further explore the effectiveness of the proposed approach, more applications on datasets of different size and sparsity are needed. Except for the optimal least squares transformations described in Section 3, optimal spline transformations can also be utilized. Splines are usually smoother and more robust, albeit at the cost of less goodness of fit with respect to the overall loss function that is minimized [12, 13]. Another direction for future work is to introduce regularization terms into the loss function of CatPCA in order to improve the stability of the algorithm.

References

1. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.T.: Application of dimensionality reduction in recommender systems - a case study. In: ACM WebKDD 2000 Web Mining for E-Commerce Workshop. (2000) 82–90
2. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval Journal **4** (2001) 133–151
3. Hofmann, T.: Latent semantic models for collaborative filtering. ACM Transactions on Information Systems **22**(1) (2004) 89–115
4. Salakhutdinov, S., Mnih, A.: Probabilistic matrix factorization. In Platt, J., Koller, D., Singer, Y., Roweis, S., eds.: Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA (2008) 1257–1264
5. Bell, M., Koren, Y.: Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In: Proceedings of 2007 Seventh IEEE International Conference on Data Mining (ICDM). (2007) 43–52
6. Tacacs, G., Pilaszy, I., Nemeth, B., Tikk, D.: Scalable collaborative filtering approaches for large recommender systems. The Journal of Machine Learning Research **10** (2009) 623–656
7. Kim, D., Yum, B.J.: Collaborative filtering based on iterative principal component analysis. Expert Systems with Applications **28**(4) (2005) 823–830
8. Paterek, A.: Improving regularized singular value decomposition for collaborative filtering. In: Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining (KDD'07), San Jose, CA, USA (2007) 39–42
9. de Leeuw, J.: Nonlinear principal component analysis and related techniques. In Greenacre, M., Blasius, J., eds.: Multiple Correspondence Analysis and Related Techniques, Chapman & Hall, Boca Raton, FL (2006) 107–133
10. Costantini, P., Linting, M., Porzio, G.: Mining performance data through nonlinear pca with optimal scaling. Applied Stochastical Models in Business and Industry **26** (2010) 85–101
11. Meulman, J., van der Kooij, A., Heiser, W.: Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In Kaplan, D., ed.: Handbook of Quantitative Methods in the Social Sciences, Newbury Park, CA: Sage Publications (2004) 49–70
12. Michailidis, G., de Leeuw, J.: The gifi system of descriptive multivariate analysis. Statistical Science **13**(4) (1998) 307–336
13. de Leeuw, J., Patrick, M.: Gifi methods for optimal scaling in r: The package homals. Journal of Statistical Software **31**(4) (2009) 1–21