Web Intelligence: Research and Development, Lecture Notes in Artificial Intelligence 2198, Springer, 394-402 (2001)

Collaborative Filtering Using Principal Component Analysis and Fuzzy Clustering

Katsuhiro Honda(1), Nobukazu Sugiura(1), Hidetomo Ichihashi(1) and Shoichi Araki(2)
(1) Graduate School of Engineering, Osaka Prefecture University,
1-1 Gakuen-cho, Sakai, Osaka, Japan
honda@ie.osakafu-u.ac.jp
http://www.ie.osakafu-u.ac.jp/~honda/

(2) Advanced Technology Research Laboratories,
Matsushita Electric Industrial Co.,Ltd.,
3-4 Hikaridai, Seika-cho, Souraku-gun, Kyoto, Japan

Abstract : Automated collaborative filtering is a popular technique for reducing information overload. In this paper, we propose a new approach for the collaborative filtering using local principal components. The new method is based on a simultaneous approach to principal component analysis and fuzzy clustering with an incomplete data set including missing values. In the simultaneous approach, we extract local principal components by using lower rank approximation of the data matrix. The missing values are predicted using the approximation of the data matrix. In numerical experiment, we apply the proposed technique to the recommendation system of background designs of stationery for word processor.

1. Introduction

Automated collaborative filtering is a popular technique for reducing information overload and has seen considerable successes in many area [1], [2], [3]. The prevalent algorithms used in the collaborative filtering are neighborhood-based methods. In the neighborhood-based methods, the subset of appropriate users is chosen based on their similarity to an active user and the weighted aggregate of their ratings is used to generate predictions for the active user. GroupLens [1], [2] first introduced an automated collaborative filtering system using a neighborhood-based algorithm and provided personalized predictions for Usenet news articles. The original GroupLens system used Pearson correlations to weight user similarity and estimate the rating by computing the weighted average of deviations from the neighbor's mean. From these points of view, the collaborative filtering can be represented as the problem of predicting missing values in a data matrix.

Missing values have frequently been encountered in data analysis in real applications. There are many approaches to handle data sets including missing values. Several methods that extract principal components without elimination or imputation of data have been proposed [4] [5] [6]. Shibayama [5] proposed a PCA(Principal Component Analysis)-like method to capture the structure of incomplete multivariate data without any imputations and statistical assumptions. The method is derived using the lower rank approximation of a data matrix including missing values, which accomplishes the minimization of the least square criterion.

In this paper, we propose a new approach for the collaborative filtering in which we estimate missing values using local principal components. The new method is based on the simultaneous application of PCA and fuzzy clustering, which is a technique for partitioning an incomplete data set including missing values into several fuzzy clusters by using local principal components. The simultaneous approaches [7] [8] to the multivariate data analysis and fuzzy clustering have been proposed since Fuzzy c-Varieties (FCV) clustering was first proposed by Bezdek *et al.* [9] [10], which can be regarded as a simultaneous approach to PCA and fuzzy clustering. FCV clustering partitions a data set into several linear clusters formed as linear varieties and thus we can extract local principal component vectors as the basis vectors of the prototypical linear varieties. Though it is difficult to describe the characteristics of a large-scale database by only one statistical model, we often obtain a practical knowledge from local model in each cluster. The least square criterion is the same as that of the objective function of FCV when no missing value is involved, hence our novel technique is an extension of FCV into incomplete data sets. By the proposed clustering, missing values in the data matrix are estimated using the local principal components. The advantage of the method is the low memory requirements. Once we obtain the local linear models, we can predict the missing values from a few simple linear models, while we need to retain all elements of the correlation matrix in the general neighborhood-based methods.

In numerical experiment, we apply the proposed technique to the recommendation system of background designs of stationery for word processor and compare the performance of the technique with that of the original GroupLens algorithm and a non-personalized prediction method.

2. Extraction of Local Principal Components and Estimation of Missing Values

Neighborhood-based methods are the most prevalent algorithms used in the collaborative filtering. In the neighborhood-based methods, the subset of appropriate users is chosen based on their similarity to an active user and the weighted aggregate of their ratings is used to generate predictions for the active user. Therefore, the collaborative filtering can be a task of predicting missing values in some user-item matrix. In this section, we introduce a technique for the extraction of local principal components and use them for the estimation of missing values.

2.1 Fuzzy c-Varieties Clustering and Local Principal Component Analysis with Least Square Criterion

The simultaneous approaches [7] [8] to the multivariate data analysis and fuzzy clustering have been proposed since Fuzzy c-Varieties (FCV) clustering was proposed by Bezdek *et al.* [9] [10], which can be regarded as a simultaneous approach to Principal Component Analysis (PCA) and fuzzy clustering. Because FCV partitions data using linear varieties as the prototypes of the clusters, we can also extract local principal component vectors as the basis of the prototypical linear varieties. Though it is difficult to describe the characteristics of a large-scale database by only one statistical model, we sometimes are able to obtain a practical knowledge from the local model in each cluster.

Let $X = (x_{ij})$ denotes a $(n \times m)$ data matrix consisting of m dimensional observation of n samples. The goal of the simultaneous approach of PCA and fuzzy clustering is to partition the data set by using local principal component vectors that represent local linear structures. FCV is a clustering method that partitions a data set into C linear clusters. The objective function of FCV with entropy regularization [11] consists of distances from data points to p dimensional prototypical linear varieties spanned by $\boldsymbol{a}_{cj}$'s as follows:

$$\begin{aligned} \min\ L_{fcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\Big\{(\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{b}_c) \\ &\quad - \sum_{j=1}^{p} \boldsymbol{a}_{cj}^{\mathrm{T}} R_{ci} \boldsymbol{a}_{cj}\Big\} + \lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \end{aligned} \tag{1}$$

$$R_{ci} = (\boldsymbol{x}_k - \boldsymbol{b}_c)(\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}}, \tag{2}$$

where u_{ci} denotes the membership degree of the data point $\boldsymbol{x}_i$ to the cth cluster and $\boldsymbol{b}_c = (b_{c1}, \cdots, b_{cm})$ is the center of the cth cluster. The entropy term of Eq.(1) is for fuzzification. The larger λ is, the fuzzier the membership assignments are. Because we derive $\boldsymbol{a}_{cj}$'s as the eigenvectors of the fuzzy scatter matrix, $\boldsymbol{a}_{cj}$'s can be regarded as local principal component vectors.

In this paper, we extract the local principal components by using least square criterion [4] [5]. We define the least square criterion for local principal component analysis using membership u_{ci} and entropy regularization as

$$\varphi = \sum_{c=1}^{C} \mathrm{tr}\Big\{(X - Y_c)^{\mathrm{T}} U_c (X - Y_c)\Big\} + \lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \tag{3}$$

where $U_c = \mathrm{diag}(u_{c1}, \cdots, u_{cn})$. $Y_c = (y_{cij})$ denotes the lower rank approximation of the data matrix X in cth cluster,

$$Y_c = F_c A_c^{\mathrm{T}} + \boldsymbol{1}_n \boldsymbol{b}_c^{\mathrm{T}}, \tag{4}$$

where $F_c = (\boldsymbol{f}_{c1}, \cdots, \boldsymbol{f}_{cn})^{\mathrm{T}}$ is the $(n \times p)$ score matrix and $A_c = (\boldsymbol{a}_{c1}, \cdots, \boldsymbol{a}_{cp})$ is the $(m \times p)$ principal component matrix. The problem is to determine F_c, A_c and $\boldsymbol{b}_c$ so that the least square criterion is minimized.

From the necessary condition $\partial\varphi/\partial\boldsymbol{b}_c = \boldsymbol{0}$ for the optimality of the objective function φ, we have

$$\boldsymbol{b}_c = (\boldsymbol{1}_n^{\mathrm{T}} U_c \boldsymbol{1}_n)^{-1} X^{\mathrm{T}} U_c \boldsymbol{1}_n, \tag{5}$$

and Eq.(3) can be transformed into

$$\begin{aligned}\varphi &= \sum_{c=1}^{C}\Big\{\mathrm{tr}(X_c^{\mathrm{T}}U_cX_c)-2\mathrm{tr}(X_c^{\mathrm{T}}U_cF_cA_c^{\mathrm{T}})\\ &\quad +\mathrm{tr}(A_cF_c^{\mathrm{T}}U_cF_cA_c^{\mathrm{T}})\Big\}+\lambda\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci},\end{aligned} \tag{6}$$

where $X_c = X - \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}}$. From $\partial\varphi/\partial F_c = O$,

$$F_cA_c^{\mathrm{T}}A_c = X_cA_c. \tag{7}$$

Under the condition that $A_c^{\mathrm{T}}A_c = I_p$, we have $F_c = X_cA_c$ and the objective function is transformed as follows:

$$\begin{aligned}\varphi &= \sum_{c=1}^{C}\Big\{\mathrm{tr}(X_c^{\mathrm{T}}U_cX_c)-\mathrm{tr}(A_c^{\mathrm{T}}X_c^{\mathrm{T}}U_cX_cA_c)\Big\}\\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci}\\ &= L_{fcv}.\end{aligned} \tag{8}$$

Therefore it can be said that Eq.(3) is equivalent to the objective function of FCV and the minimization problem is solved by computing the p largest singular values of the fuzzy scatter matrix and their associated vectors, when the data matrix doesn't include a missing value.

2.2 Local Principal Components of Data with Missing Values

Unfortunately there is no general method to deal with missing values in fuzzy clustering. Miyamoto *et al.* [12] proposed an approach that can handle missing values in Fuzzy c-Means (FCM) [10]. FCM is a fuzzy clustering method that partitions a data set into several spherical clusters. Ignoring the missing values, the objective function of FCM is written as follows:

$$\psi = \sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\sum_{j=1}^{m}d_{ij}(x_{ij}-b_{cj})^2+\lambda\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci}, \tag{9}$$

where d_{ij} is defined by

$$d_{ij} = \begin{cases} 1 & ; x_{ij} \text{ is observed.}\\ 0 & ; x_{ij} \text{ is missing.}\end{cases}$$

and the entropy term is added for fuzzification. This strategy is useful only for spherical clustering.

In this subsection, we enhance the method to partition an incomplete data set including missing values into several linear fuzzy clusters using least square criterion. To handle the missing values, we minimize only the deviations between x_{ij}'s and y_{cij}'s where x_{ij}'s are observed, and y_{cij}'s corresponding to missing values are determined incidentally.

The objective function to be minimized is defined by the convex combination of Eqs.(3), (9) and the entropy term as follows:

$$\begin{aligned}L &= \alpha\varphi+(1-\alpha)\psi+\beta\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci}\\ &= \sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\sum_{j=1}^{m}d_{ij}\Big\{\alpha(x_{ij}-\sum_{k=1}^{p}f_{cik}a_{cjk}-b_{cj})^2+(1-\alpha)(x_{ij}-b_{cj})^2\Big\}\\ &\quad +\beta\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log u_{ci},\end{aligned} \tag{10}$$

where α is a constant which defines the tradeoff between FCM and local principal component analysis. When α is 0, Eq.(10) is equivalent to Eq.(9).

To obtain a unique solution, the objective function is minimized under the constrains that

$$F_c^{\mathrm{T}}U_cF_c = I_p \quad ; \quad c = 1,\cdots,C, \tag{11}$$

$$F_c^{\mathrm{T}}\mathbf{1}_n = O \quad ; \quad c = 1,\cdots,C, \tag{12}$$

$$\sum_{c=1}^{C} u_{ci} = 1 \qquad ; \quad i = 1, \cdots, n, \tag{13}$$

and $A_c^{\mathrm{T}} A_c$ is orthogonal.

To derive the optimal A_c and $\boldsymbol{b}_c$, we rewrite Eq.(10) as follows:

$$\begin{aligned} L &= \sum_{c=1}^{C} \sum_{j=1}^{m} \Big\{ \alpha(\boldsymbol{x}_j - F_c \boldsymbol{a}_{cj} - \mathbf{1}_n b_{cj})^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - F_c \boldsymbol{a}_{cj} - \mathbf{1}_n b_{cj}) \\ &\quad + (1-\alpha)(\boldsymbol{x}_j - \mathbf{1}_n b_{cj})^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \mathbf{1}_n b_{cj}) \Big\} + \beta \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci}, \end{aligned} \tag{14}$$

where

$$\begin{aligned} X &= (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_j, \cdots, \boldsymbol{x}_m), \\ D_j &= \mathrm{diag}(d_{1j}, \cdots, d_{nj}). \end{aligned}$$

From $\partial L / \partial \boldsymbol{a}_{cj} = \mathbf{0}$ and $\partial L / \partial b_{cj} = 0$, we have

$$\boldsymbol{a}_{cj} = (F_c^{\mathrm{T}} U_c D_j F_c)^{-1} F_c^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \mathbf{1}_n b_{cj}), \tag{15}$$

$$b_{cj} = (\mathbf{1}_n^{\mathrm{T}} U_c D_j \mathbf{1}_n)^{-1} \mathbf{1}_n^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \alpha F_c \boldsymbol{a}_{cj}). \tag{16}$$

In the same way, we can derive the optimal F_c and u_{ci}. Eq.(10) is equivalent to

$$\begin{aligned} L &= \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \Big\{ \alpha(\boldsymbol{x}_i - A_c \boldsymbol{f}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}} D_i (\boldsymbol{x}_i - A_c \boldsymbol{f}_{ci} - \boldsymbol{b}_c) \\ &\quad + (1-\alpha)(\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}} D_i (\boldsymbol{x}_i - \boldsymbol{b}_c) \Big\} + \beta \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci}, \end{aligned} \tag{17}$$

and $\partial L / \partial \boldsymbol{f}_{ci} = \mathbf{0}$ and $\partial L / \partial u_{ci} = 0$ yields

$$\boldsymbol{f}_{ci} = (A_c^{\mathrm{T}} D_i A_c)^{-1} A_c^{\mathrm{T}} D_i (\boldsymbol{x}_i - \boldsymbol{b}_c), \tag{18}$$

$$\begin{aligned} u_{ci} &= \exp\Big\{ -\Big(\alpha(\boldsymbol{x}_i - A_c \boldsymbol{f}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}} D_i (\boldsymbol{x}_i - A_c \boldsymbol{f}_{ci} - \boldsymbol{b}_c) \\ &\quad + (1-\alpha)(\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}} D_i (\boldsymbol{x}_i - \boldsymbol{b}_c) \Big) / \beta - 1 \Big\}, \end{aligned} \tag{19}$$

where

$$\begin{aligned} X &= (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_n)^{\mathrm{T}}, \\ D_i &= \mathrm{diag}(d_{i1}, \cdots, d_{im}). \end{aligned}$$

The proposed algorithm can be written as follows.

Step1 Initialize $U_c, A_c, \boldsymbol{b}_c, F_c$ randomly in each cluster and normalize them so that they satisfy the constraints Eqs.(11)-(13) and $A_c^{\mathrm{T}} A_c$ is orthogonal.

Step2 Update A_c's using Eq.(15) and transform them so that each $A_c^{\mathrm{T}} A_c$ is orthogonal.

Step3 Update F_c's using Eq.(18) and normalize them so that they satisfy the constraints Eqs.(11) and (12).

Step4 Update $\boldsymbol{b}_c$'s using Eq.(16).

Step5 Update U_c's using Eq.(19) and normalize them so that Eq.(13) holds.

Step6 If

$$\max_{i.c} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \epsilon,$$

then stop. Otherwise, return to Step3.

2.3 Estimation of Missing Values Using Local Principal Components

We propose a technique for estimating missing values using the lower rank approximations of the data matrix derived in local principal component analysis. Because the lower rank matrix Y_c derived in each cluster includes no missing values, we can estimate the missing values in the data matrix X by replacing them with the corresponding elements of Y_c. It means that we estimate the missing values on the assumption that the data points including missing values are on the linear varieties spanned by local principal component vectors. The procedure entails two steps.

1. Select the cluster c in which the active user i has largest membership u_{ic}.

2. Predict the missing value x_{ij} from the corresponding element y_{cij}.

The clustering part and the lower rank approximation part correspond to the selection of neighbors and the prediction using selected neighbors' ratings in the neighborhood-based methods respectively. Therefore, it can be said that the novel technique is one of the neighborhood-based methods where the selection of neighbors and the prediction of missing values are applied simultaneously.

Once we estimate the local linear models, we can predict the ratings of new active users. The memberships and the principal component scores of the new active users are estimated by Eqs.(19), (18) and we can predict y_{cij} using Eq.(20).

$$y_{cij} = \sum_{k=1}^{p} f_{cik} a_{cjk} + b_{cj} \tag{20}$$

3. Experimental Results

We implemented the novel technique presented in the previous section for the collaborative filtering and tested them with "kansei" data set that is the set of psychological evaluation data composed of 285 instances in which each user evaluated 8 different images used as background designs of stationery for word processor. Fig. 1 shows the 8 background designs. Each user evaluated the images on a scale from 1 to 7 based on the

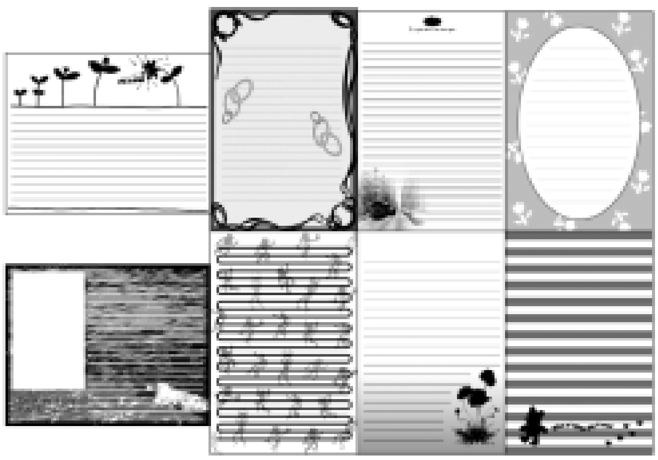

Figure 1: 8 background designs

semantic differential (SD) method as shown in Fig.2.

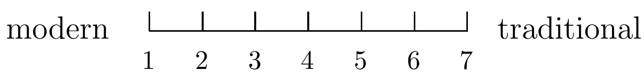

Figure 2: Seven-point Rating Scale

Each user withheld ratings for 2 items randomly and the predicted ratings were computed for those 2 items using local principal components derived by the proposed clustering method. The tradeoff parameter α and the weighting parameter of fuzziness β were set to 0.8 and 0.5 respectively, and the users were partitioned into two clusters. In addition, we also predicted the ratings using original GroupLens [1] and a non-personalized prediction method [13]. In the non-personalized prediction method, we computed the ratings using deviation-from-mean average over all users. Table 1 compares the result of them.

Table 1: Comparison of Results

Algorithm	**MAE**	**ROC**
Non-personalized Method	0.963	0.479
Original GroupLens	1.055	0.513
Proposed Method	1.066	0.591

For assessing the accuracy of the three prediction methods, we used not only the mean absolute error (MAE), but also the receiver operating characteristic (ROC) sensitivity. ROC sensitivity is a measure of the diagnostic power of a filtering system [13]. The sensitivity refers to the probability of a randomly selected good item being accepted by the filter. The greater the value is, the richer the performance becomes. The maximum value is one. In this paper, the items whose ratings are larger than 5 or smaller than 3 are regarded as good items. Although the proposed method had largest MAE, it provides the best performance according to ROC sensitivity. It indicates that the proposed method possesses the ability to recommend good items to active users.

4. Conclusions

In this paper, we proposed a new approach to the collaborative filtering. The new method is based on a simultaneous application of PCA and fuzzy clustering and is a kind of neighborhood-based methods. For the prediction of the ratings, users are partitioned into several linear clusters. It can be said that the clustering part is responsible for the selection of the neighborhood. The larger the cluster number is, the fewer the neighbors are.

Our future work is to determine the ability of the proposed method by using some other benchmarks.

References

[1] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: An Open Architecture for Collaborative Filtering of Netnews. Proc. of ACM Conference on Computer-Supported Cooperative Work (1994) 175–186

[2] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gardon, L. R., Riedl, J.: Grouplens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, Vol.40, No.3 (1997) 77–87

[3] Shardanand, U., Maes, P.: Social Information Filtering: Algorithms for Automating "Word of Mouth". Proc. of ACM Conference on Human Factors in Computing Systems (1995) 210–217

[4] Wiberg, T.: Computation of Principal Components when Data are Missing. Proc. of 2nd Symposium on computational Statistics (1976) 229–236

[5] Shibayama, T.: A PCA-Like Method for Multivariate Data with Missing Values. Japanese Journal of Educational Psychology, Vol.40 (1992) 257–265 (in Japanese)

[6] Shum, H., Ikeuchi, K., Reddy,R.: Principal Component Analysis with Missing Data and its Application to Polyhedral Object Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 9 (1995) 854–867

[7] Yamakawa, A., Honda, K., Ichihashi, H., Miyoshi, T.: Simultaneous Approach to Fuzzy Cluster, Principal Component and Multiple Regression Analysis. Proc. of International Conference on Neural Networks (1999)

[8] Oh, C.-H., Komatsu, H., Honda, K., Ichihashi, H.: Fuzzy Clustering Algorithm Extracting Principal Components Independent of Subsidiary Variables. Proc. of International Conference on Neural Networks (2000)

[9] Bezdek, J. C., Coray, C., Gunderson, R., Watson, J.: Detection and Characterization of Cluster Substructure 2. Fuzzy c-Varieties and Convex Combinations Thereof. SIAM J. Appl. Math., Vol.40, No.2 (1981) 358–372

[10] Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)

[11] Miyamoto, S., Mukaidono, M.:Fuzzy *c*-Means as a Regularization and Maximum Entropy Approach. Proc. of 7th International Fuzzy Systems Association World Congress, Vol.2 (1997) 86–92

[12] Miyamoto, S., Takata, O., Umayahara, K.: Handling Missing Values in Fuzzy c-Means. Proc. of 3rd Asian Fuzzy Systems Symposium (1998) 139–142

[13] Herlocker, J. L., Konstan, J. A., Borchers, A. Riedl, J.: An Algorithmic Framework for Performing Collaborative Filtering. Proc. of Conference on Research and Development in Information Retrieval (1999)