Weighted consensus clustering for multiblock data

Ndèye Niang, Mory Ouattara

Weighted consensus clustering for multiblock data

Ndèye Niang*, Mory Ouattara**

*Statistique Appliquée, CNAM 292, rue Saint Martin, 75141 Paris Cedex 03, France
**Laboratoire de Mathématique Informatique,
UFR-SFA, Université Nangui Abrogoua, 21 BP 1288 Abidjan 21,
ouattaramory.sfa@univ-na.ci

Résumé. Nous nous intéressons au problème de classification d'individus décrits par plusieurs variables structurées en blocs homogènes. Nous le formulons comme la recherche de consensus de partitions et nous proposons une méthode de consensus pondéré de partitions basée sur le coefficient RV de corrélation vectorielle. L'idée principale de la méthode proposée consiste à agréger les partitions initiales séparées obtenues à partir de chaque bloc en une partition globale la plus similaire à ces partitions initiales au sens du coefficient RV. La méthode proposée est comparée à la méthode CSPA (Cluster based Similarity Partitioning Algorithm) et une méthode de consensus pondéré basée sur la factorisation non négative de matrices.

1 Introduction

The general problem addressed in this paper is clustering individuals described by variables which are divided in several homogeneous and meaningful blocks. Since blocks are assumed to be homogeneous, preserving this block homogeneity would help to exhibit the underlying structure of individuals. Thus in a first level, the individuals are clustered according to each block separately and the resulting partitions (called contributory partitions) are aggregated into a consensus partition in a second step. Therefore, the issue of this paper is reformulated as a problem of consensus of partitions. The choice of the first step clustering method is not addressed here. We only focus on the aggregation of the obtained partitions.

Clustering multiblock data has been addressed by several consensus methods. A survey on these methods can be found in Day (1994). The main idea of consensus methods is to agglomerate the separate partitions obtained from each block into a global partition that must be as similar as possible to the contributory partitions according to some index, eg. the Rand index. Other more recent methods in the ensemble clustering approach (Strehl, 2003) have also addressed the consensus clustering problem. The contributory partitions are seen as categorical variables and then are associated with their indicator matrices and connectivity matrices whose entries are 1 if two individuals are in the same cluster and 0 otherwise. (Strehl, 2003) proposed CSPA (Cluster based Similarity Partitioning Algorithm) which consists of re-clustering the individuals using a so-called association matrix considered as a similarity matrix. The association matrix is obtained by simply averaging the connectivity matrices. Thus its entries are defined as the fraction of partitions in which two individuals are in the same cluster. In this method,

implicitly, all contributory partitions are treated equally, despite the facts that (1) partitions could differ significantly and (2) subsets of partitions could be highly correlated. Therefore as pointed out in (Tao, 2008), this approach essentially based on a simple averaging process is inadequate. CSPA yields unstable and biased results. Tao (2008) proposed a weighted consensus clustering method (denoted WNMF) that do not have the CSPA drawbacks. It is based on a non-negative matrix factorization (NMF) and the optimization of a quadratic function to obtain the consensus partition and the weights respectively.

We propose a weighted consensus clustering method based on the RV correlation coefficient (Robert, 1976) between the connectivity matrices to find an unique partition of individuals from contributory partitions. A so-called compromise matrix, weighted average of connectivity matrices, is used to re-cluster the individuals with a classical hierarchical algorithm. In the next section we detail our method which can be seen as the combination of WNMF in the sense that we obtain a weighted consensus matrix and of CSPA as we cluster this later compromise matrix.

2 Weighted consensus clustering

2.1 RV coefficient

We consider P variables divided into T blocks. In each block N outcomes of p_t variables are available. The RV index Robert (1976) is a measure of the relationship between two sets of variables X_t and $X_{t'}$. Prior to the analysis, data matrices are usually centered and normalized to remove scale effects. X_t is associated with the matrix of scalar products $W_t = X_t X_t'$, where X_t' denotes the transpose of X_t. This step is necessary to have square matrices of same dimension N and then allows to have X_t matrices with different number of columns. The RV between X_t and $X_{t'}$ derived from Hilbert-Schmidt's scalar product, is defined as follows :

$$RV(W_t, W_t') = \frac{trace(W_t W_{t'})}{\sqrt{trace(W_t^2)trace(W_{t'}^2)}}$$

RV index is non-negative and scaled between 0 and 1. The closer the RV index is to 1, the more similar the matrices X_t and $X_{t'}$ are.

2.2 RV based consensus clustering

Let P_t be the t^{th} contributory partition obtained from the clustering of the t^{th} block of variables into K_t clusters (which may differ from one partition P_t to another). P^* denotes the final partition, consensus of the T partitions P_t. We consider $X_t,\ t = 1, \ldots, T$ the indicator matrix of K_t dummy variables related to the categorical variable associated with the t^{th} contributory partition. Each of these matrices is associated with a matrix $W_t = X_t X_t'$ which is a connectivity matrix whose entries are : $W_t(i,j) = \begin{cases} 1 & if \quad P_t(i) = P_t(j) \\ 0 & otherwise \end{cases}$

The first step of the proposed weighted consensus clustering method is to find a so-called compromise matrix W^*, weighted average of the W_t which has to be the most similar to the W_t (representing the contributory partitions). Our proposition is to use the RV index as a

measure of this similarity between data tables. Therefore the criterion to optimize to get the compromise matrix is the following one :

$$\underset{\alpha_1,\ldots,\alpha_T}{max} \sum_{t=1}^{T} RV(W_t, W^*) = \underset{\alpha_1,\ldots,\alpha_T}{max} \sum_{t=1}^{T} RV(W_t, \sum_{t=1}^{T} \alpha_t W_t)$$

We are looking for the best linear combination of the matrices W_t that maximizes its vectorial correlation RV with the W_t matrices. As in principal component analysis, the solution is obtained through weights equal to the coordinates of the first standardized eigenvector α^1 of the $T \times T$ square matrix S whose elements are RV coefficients between every pair of indicator matrices (Lavit, 1984). Since all elements of S are non-negative, the coordinates α_t^1 are all non-negative and used to get the consensus matrix W^* between the T connectivity matrices : $W^* = \sum_{t=1}^{T} \alpha_t^1 W_t$. The weights α_t^1 represent the agreement between data tables and the compromise matrix.
We propose to apply hierarchical ascendant clustering algorithm on this compromise matrix considered as a similarity matrix to re-cluster the individuals as in CSPA. Therefore our method, denoted RVCONS, can be seen as a combination of WNMF considering the weighted aggregate matrix and of CPSA as we cluster the compromise matrix using classical hierarchical algorithm (rather than NMF). However, there are differences between the sets of weights associated with WNMF and RVCONS. Firstly, the way to get these weights is different : WNMF uses a iterative algorithm while RVCONS weights come from eigen elements derivations or singular value decomposition. Secondly, WNMF removes redundancy by giving an important weight to one single partition among highly correlated ones but in the same time it would give a important weight to a partition very different from the others. At the opposite, with RVCONS, this later partition would be considered as an outlier and then associated with a small weight.

3 Applications

We exemplify the proposed method on a simulated data set and on one from the UCI repository. We compare our consensus clustering method RVCONS with CSPA and WNMF. We use data sets with labelled individuals in order to have a reference partition.

3.1 Data sets descriptions

The simulated data set denoted D1, consists of 11 blocks of 5 variables. The RV coefficients between pair of blocks among the first 10 blocks are about 0.80, that is the blocks are highly correlated. We add one more block with a lower RV coefficient equal to 0.01 with each of the first 10 blocks.
The UCI Multiple Features data set contains 2000 individuals of handwritten numerals that were extracted from a collection of Dutch utility maps. These patterns were classified into 10 clusters ("0"–"9"), each having 200 individuals. Each individual is described by 649 features divided into the following 6 feature groups denoted B_1, B_2, B_3, B_4, B_5 and B_6 :

— mfeat-fou block : contains 76 Fourier coefficients of the character shapes ;
— mfeat-fac block : contains 216 profile correlations ;
— mfeat-kar block : contains 64 Karhunen-Loeve coefficients ;

	B_1	B_2	B_3	B_4	B_5	B_6
B_1	1	0.985	0.706	0.312	0.967	0.987
B_2	0.985	1	0.678	0.433	0.935	0.965
B_3	0.706	0.678	1	0.199	0.764	0.736
B_4	0.312	0.433	0.199	1	0.273	0.272
B_5	0.967	0.935	0.764	0.273	1	0.979
B_6	0.987	0.965	0.736	0.272	0.979	1

TAB. 1 – *the RV coefficients between the 6 blocks*

— mfeat-pix block : contains 240 pixel averages in 2D 3 windows ;
— mfeat-zer block : contains 47 Zernike moments ;
— mfeat-mor block : contains six morphological features.

Table 1 contains the RV coefficients between these 6 blocks. In this exemple there are different levels of similarity between blocks : blocks 1, 2, 5 and 6 are very highly correlated, block 3 has quite high RV values while block 4 has the lowest RV values and can be considered as not similar to the others blocks.

3.2 Application of consensus methods

The consensus methods have been applied with 30 initialisations of K-means to obtain initial partitions. The results in table 2 for the simulated data show quite similar initial partitions for the first 10 blocks with high accuracy mean values. The partition associated with the last block has lower accuracy mean values (around 0.5) for others contributory partitions as well as the reference partition. In table 3 are the means of 30 accuracy and Adjusted Rand indices computed between the consensus and contributory partitions. Table 4 contains the means of the weights used to get the compromise matrix.

Table 3 shows that RVCONS performs better than WNMF for the contributory partitions (excepted for P_{10}) as well as the reference partition. Moreover, comparing the last columns of table 2 and table 3, it can be seen that RVCONS consensus partition improves the accuracy and Adjusted Rand indices for the reference partition. RVCONS provides equal weights for the 10 highly correlated blocks and assigns a quite zero weight to the noisy block. This can explain the similar performances of RVCONS and CSPA. At the opposite, WNMF gives the most important weight to the last noisy block which has the lowest similarity with the reference partition. We consider this as a drawback of the WNMF method.
For the DMU data set, table 6(a), shows that the 3 consensus partitions have quite similar accuracy and Adjusted Rand indices for the reference partition. But there are differences for the contributory partitions : P_1, P_3 and P_5 have greater WNMF accuracy or Adjusted Rand indices than the RVCONS ones. RVCONS provides weights for the 6 blocks according more or less to their level of accuracy with the reference partition while WNMF gives the most important weight to the block 1 (associated with one of the lowest accuracy level with the reference partition) and the block 3 (associated with one with the highest accuracy with the reference partition). This may be related to LASSO constraints on WNMF weights (Tao, 2008).

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	Label
P_1	1.00	0.83	0.85	0.78	0.73	0.77	0.82	0.80	0.82	0.74	0.51	0.76
P_2	0.83	1.00	0.92	0.84	0.78	0.86	0.90	0.88	0.89	0.83	0.51	0.85
P_3	0.85	0.92	1.00	0.87	0.79	0.88	0.91	0.90	0.91	0.84	0.53	0.86
P_4	0.78	0.84	0.87	1.00	0.78	0.83	0.85	0.85	0.85	0.78	0.51	0.79
P_5	0.73	0.78	0.79	0.78	1.00	0.76	0.80	0.77	0.78	0.73	0.53	0.81
P_6	0.77	0.86	0.88	0.83	0.76	1.00	0.86	0.86	0.85	0.81	0.52	0.81
P_7	0.82	0.90	0.91	0.85	0.80	0.86	1.00	0.87	0.88	0.84	0.53	0.85
P_8	0.80	0.88	0.90	0.85	0.77	0.86	0.87	1.00	0.87	0.81	0.50	0.83
P_9	0.82	0.89	0.91	0.85	0.78	0.85	0.88	0.87	1.00	0.82	0.51	0.84
P_{10}	0.74	0.83	0.84	0.78	0.73	0.81	0.84	0.81	0.82	1.00	0.51	0.80
P_{11}	0.51	0.51	0.53	0.51	0.53	0.52	0.53	0.50	0.51	0.51	1.00	0.52

TAB. 2 – *The mean of the Accuracy between the initial partitions and the reference partition for D1 data set*

Method	Index	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	Label
CSPA	Acc	0.89	0.97	0.94	0.92	0.88	0.92	0.93	0.95	0.94	0.88	0.62	0.88
	AR	0.91	0.93	0.93	0.92	0.89	0.91	0.89	0.93	0.93	0.86	0.22	0.83
WNMF	Acc	0.80	0.84	0.83	0.80	0.77	0.81	0.83	0.83	0.82	0.79	0.70	0.85
	AR	0.89	0.90	0.90	0.89	0.86	0.88	0.86	0.90	0.90	0.88	0.25	0.80
RVCONS	Acc	0.89	0.97	0.94	0.91	0.88	0.92	0.96	0.95	0.94	0.87	0.62	0.88
	AR	0.92	0.93	0.94	0.92	0.89	0.91	0.89	0.93	0.93	0.86	0.22	0.83

TAB. 3 – *Accuracy and Adjusted Rand index between the initial partitions, the reference partition and the consensus partition for D1 data set*

Table	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}	W_{11}
RVCONS	0.10	0.10	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10	**0.04**
WNMF	**0.16**	0.08	0.06	**0.12**	0.06	0.05	0.09	0.09	**0.14**	**0.15**	**0.41**

TAB. 4 – *Compromise coefficients for D1 data set*

	P_1	P_2	P_3	P_4	P_5	P_6	Label
P_1	1.00	0.72	0.57	0.69	0.46	0.49	0.64
P_2	0.72	1.00	0.71	0.84	0.53	0.53	0.86
P_3	0.57	0.71	1.00	0.66	0.46	0.40	0.70
P_4	0.69	0.84	0.66	1.00	0.53	0.50	0.81
P_5	0.46	0.53	0.46	0.53	1.00	0.46	0.57
P_6	0.49	0.53	0.40	0.50	0.46	1.00	0.51

TAB. 5 – *The mean of the Accuracy between the initial partitions and the reference partition for DMU data set*

4 Conclusion

We presented a method for clustering multiblock data based on the RV index and a simple eigenvector derivation to define a consensus similarity matrix which is used to re-cluster the individuals to find a consensus partition. The first results of its application on simulated data as

(a) Accuracy and Adjusted Rand Index between the consensus partition and the initial partitions and the reference partition

Method	Index	P_1	P_2	P_3	P_4	P_5	P_6	Label
CSPA	Acc	0.87	0.93	0.90	0.94	0.86	0.85	0.81
	AR	0.46	0.55	0.71	0.71	0.35	0.31	0.71
WNMF	Acc	0.87	0.91	0.92	0.91	0.86	0.83	0.80
	AR	0.47	0.40	0.65	0.55	0.47	0.31	0.70
RVCONS	Acc	0.86	0.93	0.90	0.94	0.86	0.84	0.80
	AR	0.45	0.56	0.74	0.74	0.35	0.31	0.70

(b) Compromise coefficients

Table	W1	W2	W3	W4	W5	W6
RVCONS	0.15	0.19	0.17	0.19	0.16	0.15
WNMF	0.24	0.12	0.26	0.14	0.14	0.10

TAB. 6 – *The mean of the Accuracy, Adjusted Rand Index and the weights of WNMF and RVCONS for the DMU data set*

well as one real data show better performances compare to WNMF, another weighted consensus clustering method. Work is on progress for more formal evaluations on simulated data as well as real one from batch process monitoring. In addition, we still studying the weights assignment step particulary in case of large number of blocks considering sparsity.

Références

Day, W.E. (1994). *Foreword: Comparison and consensus of classifications.*. Journal of Classification, 3(2), pp. 183-185.

Lavit, C., Escoufier,Y., Sabatier, R., et P. Traissac (1984). *The act (statis method)*. Statistics & Data Analysis. 18(1), 97)-119

Strehl, A., et Ghosh, J (2003). *Cluster ensembles a knowledge reuse framework for combining multiple partitions*. The Journal of Machine Learning Research. 3, 583-617

Tao, L. et Ding, C (2008). *Weighted consensus clustering*. Mij. 1(2), 97-119 (2008)

Robert, P. et Escoufier, Y (1976). *A unifying tool for linear multivariate statistical methods: the RV-coefficient*. Applied statistics. 27(3), 257-265

Summary

We address the issue of clustering individuals described by several homogeneous blocks of variables. Reformulating it as a problem of consensus of partitions, we propose a weighted consensus method based on the RV index to find the partition of the individuals. The main idea of this consensus method is to agglomerate the separate initial partitions obtain from each block into a global partition which has to be the most similar to the initial partitions according to the RV coefficient. The proposed method is illustrated and compared to CSPA (Cluster based Similarity Partitioning Algorithm) and a weighted consensus clustering method based on non-negative matrix factorization.