Low-Rank Joint Subspace Construction for Multimodal Data Clustering

Aparajita Khan
March 14, 2019
Multimodal Data Clustering

Common set of $n$ samples
$M$ sets of observations
Multimodal Data Clustering

Common set of $n$ samples
$M$ sets of observations

$X_1 \in \mathbb{R}^{n \times d_1}$           $X_2 \in \mathbb{R}^{n \times d_2}$        $\ldots$               $X_M \in \mathbb{R}^{n \times d_M}$

$n << d_m$, for $m = 1, \ldots, M$
Multimodal Data Clustering

Common set of $n$ samples
$M$ sets of observations

$X_1 \in \mathbb{R}^{n \times d_1}$ \hspace{1cm} $X_2 \in \mathbb{R}^{n \times d_2}$ \hspace{1cm} \ldots \hspace{1cm} $X_M \in \mathbb{R}^{n \times d_M}$

$n << d_m$, for $m = 1, \ldots, M$

Cluster the $n$ samples considering information from $M$ different modalities/views
Multimodal Data Clustering

\[ X_1 \in \mathbb{R}^{n \times d_1} \quad X_2 \in \mathbb{R}^{n \times d_2} \quad \ldots \quad X_M \in \mathbb{R}^{n \times d_M} \]

Lower dimensional subspace

Clustering
Joint Subspace construction

Statistical hypothesis testing based Rank Estimation

\[ X_m = \Xi_m + Z_m, \quad \text{signal} \ \Xi_m \sim \sum_{j=1}^{k} \pi_j \mathcal{N}(\mu_j, \Sigma_j), \quad \text{noise} \ Z_m \sim \mathcal{N}(0, \Lambda) \]
Joint Subspace construction

Subspaces $U_1^r, U_2^r, \ldots, U_M^r$

Principal Subspace of $X_m$

$X_m = U_m \Sigma_m V_m^T$

Select Only Relevant Cluster Information from Each Subspaces
Relevance: 
\[ R_l(X_m) = \frac{1}{2} \left[ 1 + \frac{H_{r_m} - F_{\chi^2}^{-1}((1 - \alpha), r_m)}{\max\{H_{r_m}, F_{\chi^2}^{-1}((1 - \alpha), r_m)\}} \right] \]

\( \alpha = 0.05 \)
\( \tau_m = F_{\chi^2}^{-1}(0.95, r_m) \)
\( \delta_m = H_{r_m} - \tau_m \)
\( r_1 \neq r_2 \neq r_3 \)

Royston’s H-statistic for \( r_m \) dimensional principal subspace of modality \( X_m \)
Joint Subspace construction

subspaces $U_1^r, U_2^r, \ldots, U_M^r$

CASE 1: Residuals follow normal distribution

CASE 2: Residuals deviate from normality

joint basis

$U_2 = [U_1^r \ b_2^r]$
Joint Subspace construction

subspaces $U_1^r, U_2^r, \ldots, U_M^r$

joint basis
$U_2 = [U_1^r \quad \chi_2^r]$

append

compute joint basis $U_M$ from $M$ subspaces
Joint Subspace construction

subspaces $U_1^r, U_2^r, \ldots, U_M^r$

joint basis

$U_2 = [U_1^r \ \chi_2^r]$

append

projection

M subspaces

compute joint basis $U_M^r$ from $M$ subspaces

Perform clustering in Joint Subspace
Data sets:
Colorectal carcinoma (CRC) [307 samples, 2 clusters]
Cervical carcinoma (CESC) [124 samples, 3 clusters]
Lower grade glioma (LGG) [267 samples, 3 clusters]
Stomach adenocarcinoma (STAD) [199 samples, 4 clusters]

Modalities: DNA Methylation
Gene Expression
Micro-RNA Expression
Protein Expression
Importance of Noise-Free Approximation

- Performance of selected proposed joint subspace is better than taking full-rank subspace from each modality
Proposed method performs better than all the individual modalities in all 4 data sets

Very high performance for LGG and STAD data sets
Results

Comparison with Existing Approaches

- Performance of proposed better than existing in all data sets, especially for STAD data set
- PCA and SNF have next best performance
Conclusion

- Statistical hypothesis testing to estimate rank of subspaces
- Multivariate Normality based relevance of individual modalities
- Select only relevant modalities
- Select only complementary information between two modalities

A. Khan and P. Maji, "Low-Rank Joint Subspace Construction for Cancer Subtype Discovery," IEEE/ACM Transactions on Computational Biology and Bioinformatics. (Accepted)


THANK YOU!