Subspace Clustering With Application To Text Data

Hankui Peng[1], Nicos Pavlidis[1], Idris Eckley[1], Ioannis Tsalamanis[2]

Lancaster University (LU)[1], Office for National Statistics (ONS)[2]

Motivation

- The Office for National Statistics (ONS) are experimenting with incorporating web-scraped data into the price index generating process.
- Clustering methods could be used to automate this process effectively and efficiently.

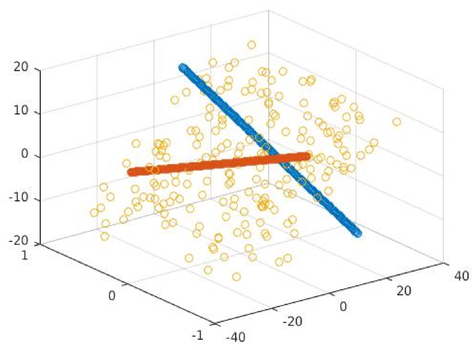
Introduction

- We study the problem of grouping short text data using subspace clustering.
- The problem of clustering short text data arises in many application domains, such as sentiment analysis, product categorisation.
- One challenge of such tasks comes from the fact that their vectorial representation are usually high-dimensional.
- Additionally, the text lengths are generally short for online products.

Subspace Clustering

It refers to the problem of separating data according to their underlying subspaces [1],

$$S_k = \left\{ \mathbf{x} \in \mathbb{R}^P : \mathbf{x} = U_k \mathbf{y}, k = 1, \ldots, K \right\}. \quad (1)$$



- $U_k \in \mathbb{R}^{P \times d_k}$ is the set of basis vectors for subspace S_k.
- $\mathbf{y} \in \mathbb{R}^{d_k \times 1}$ is a low dimensional representation for the original data object $\mathbf{x}$.

Reduced Row Echelon Form

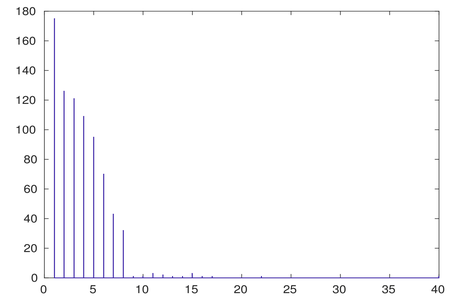
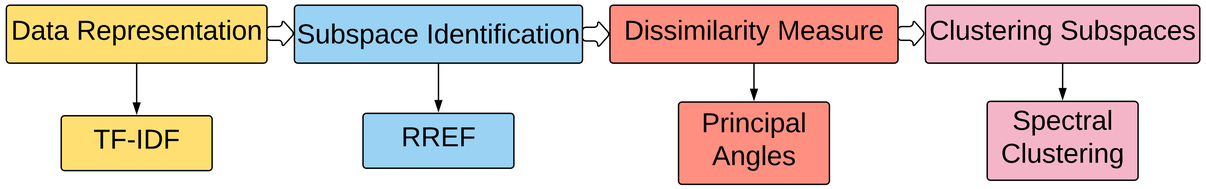


Figure 1:Histogram of the number of observations in each subspace identified through RREF.

- TF-IDF representation is used to transform text data into vectorial representation for clustering.



Principal Angles

- Let $\mathcal{S}_i$ and $\mathcal{S}_j$ be two subspaces with $1 \leqslant \dim \mathcal{S}_i = d_i \leqslant \dim \mathcal{S}_j = d_j$; $Q_{\mathcal{S}_i}$, $Q_{\mathcal{S}_j}$ the matrices of orthonormal basis vectors.
- The *principal angles*, $0 \leqslant \theta_1 \leqslant \theta_2 \leqslant \ldots \leqslant \theta_{d_i} \leqslant \pi/2$, between $\mathcal{S}_i$ and $\mathcal{S}_j$ can be defined recursively for $k = 1, \ldots, d_i$ as,

$$\cos(\theta_k) = \max_{u \in \mathcal{S}_i} \max_{v \in \mathcal{S}_j} \cos(u^\top v) = u_k^\top v_k,$$

subject to

$$\|u\| = \|v\| = 1, \text{ and}$$
$$u^\top u_m = 0, \; v^\top v_m = 0, \text{ for } 0 < m < k.$$

- Principal angles between any pair of subspaces $\mathcal{S}_i$ and $\mathcal{S}_j$ can be obtained through SVD,

$$Q_{\mathcal{S}_i}^\top Q_{\mathcal{S}_j} = Y \Sigma Z^\top,$$
$$\theta_k = \arccos(\Sigma(k,k)), \; i \in \{1, \ldots, d_i\}.$$

Minimum Angle Clustering

1. Transform the TF-IDF matrix into its reduced row echelon form X_{rref}.
2. Define matrix Y to capture the pairwise connectivity through $Y(i,j) = \mathbf{1}(X_{\text{rref}}(i,j) \neq 0)$.
3. Construct a graph $G(A)$ based on adjacency matrix $A = Y^T Y$ and obtain the set of connected components $\{c_1, c_2, \ldots, c_{n_c}\}$.
4. Construct a dissimilarity matrix D using the proposed measure,

$$\begin{aligned} \mathrm{D}(i,j) &= \frac{1}{d_j}\left(d_j - d_i + \sum_{i=1}^{d_i}(1 - \cos(\theta_i))\right) \\ &= 1 - \frac{1}{d_j}\sum_{i=1}^{d_i}\cos(\theta_i). \end{aligned}$$

5. Apply spectral clustering to D.

Results

- The TF-IDF data matrix contains 2921 data objects and 2106 features.
- Spectral clustering with both the TF-IDF matrix X and the adjacency matrix A are conducted.
- Comparisons are made to both subspace and non-subspace methods.

Method	MAC	SSC	LRR	PKM
Purity	**0.742**	0.219	0.510	0.591
NMI	**0.328**	0.032	0.041	0.218
ARI	**0.251**	0.025	-0.023	0.191
Method	**SC(X)**	**SC(A)**	**LDA**	**PDDP**
Purity	0.512	0.519	0.510	0.578
NMI	0.022	0.052	0.021	0.084
ARI	0.000	-0.024	0.011	0.065

Table 1:Performance comparison on Amazon dataset.

Conclusions

We propose a novel subspace clustering technique called *Minimum Angle Clustering (MAC)* that models data from different clusters as distinct subspaces [2].

- It first identifies low-dimensional subspaces that contain small clusters of texts.
- To merge these into meaningful clusters we utilise principal angles to quantify the dissimilarity between linear subspaces.
- Experimental results on an Amazon product names dataset show that MAC compares favourably with standard and subspace methods.

Future Directions

In future work, we wish to investigate the following directions:

- Exploit the hierarchical structure for suitable applications using text data.
- Incorporate active learning approaches to inform and guide the clustering procedure.

References

[1] René Vidal.
Subspace clustering.
IEEE Signal Processing Magazine, 28(2):52–68, 2011.

[2] Hankui Peng, Nicos Pavlidis, Idris Eckley, and Ioannis Tsalamanis.
Subspace clustering of very sparse high-dimensional data.
Proceedings of 2018 IEEE International Conference on Big Data (To appear), 2018.

Acknowledgements

This research has made use of The Billion Prices Project Dataverse, which is maintained by Harvard University & MIT.

STOR-i excellence with impact

Office for National Statistics