Co-clustering for Data Science

Extended Abstract†

Mohamed Nadif
LIPADE, University of Paris Descartes
45, rue des Saints Pères, 75006 Paris
France
mohamed.nadif@parisdescartes.fr

ABSTRACT

Many of the datasets encountered in statistics are two-dimensional in nature and can be represented by a matrix. Classical clustering procedures seek to construct separately an optimal partition of rows or, sometimes, of columns. In contrast, co-clustering methods cluster the rows and the columns simultaneously and organize the data into homogeneous blocks (after suitable permutations). Methods of this kind have practical importance in a wide variety of applications such as document clustering, bioinformatics, and collaborative filtering. Our goal is to present a comprehensive survey of co-clustering –models and algorithms– under different approaches.

KEYWORDS

Co-clustering, Data Science

Biography

Co-clustering or block clustering [8-13, 4] is an important extension of traditional one-sided clustering that addresses the problem of simultaneous clustering of both dimensions of data matrices. Since the works of [16, 5], co-clustering, under various names, has been successfully used in a wide range of application domains where the co-clusters can take different forms. For instance, in bioinformatics co-clustering, referred to as biclustering [19, 14, 15], is used to cluster genes and experimental conditions simultaneously, in collaborative filtering [17] to group users and items simultaneously, and in text mining [1-3, 6, 7, 18] to group terms and documents simultaneously.

Co-clustering exhibits several practical advantages making it possible to meet the growing needs in several current areas of interest, in terms of effectiveness, scalability and visualization. Below, we summarize some key properties of co-clustering:

- By intertwining row clustering and column clustering at each stage, co-clustering performs an implicitly adaptive dimensionality reduction, which is imperative to deal with high dimensional sparse data. This makes it possible (i) to develop efficient algorithms with a dramatically smaller number of parameters (ii) to reduce the original data matrix into a much simpler and condensed data matrix with the same structure.
- Co-clustering exploits the inherent duality between rows and columns of data matrices making it possible to enhance the clustering along both dimensions, by using the information contained in column clusters during row assignments and vice versa.
- Far from adding complexity, co-clustering is more informative than one-sided clustering, and produces meaningful clusters. In the case of document-term matrices, for example, co-clustering annotates sets of documents automatically by clusters of terms.

Several approaches have been proposed in order to address the problem of co-clustering and to date, there is no co-clustering approach that works better than the others in all situations. We aim to give a comprehensive survey of co-clustering.

REFERENCES

[1] Ailem M, Role F, Nadif M (2016) Graph modularity maximization as an effective method for co-clustering text data. Knowledge-Based Systems 109:160{173

[2] Ailem M, Role F, Nadif M (2017a) Model-based co-clustering for the effective handling of sparse data. Pattern Recognition 72:108-122

[3] Ailem M, Role F, Nadif M (2017b) Sparse poisson latent block model for document clustering. IEEE Transactions on Knowledge and Data Engineering 29(7):1563-1576

[4] Banerjee A, Dhillon IS, Ghosh J, Merugu S, Modha DS (2007) A generalized maximum entropy approach to bregman co-clustering and matrix approximations. Journal of Machine Learning Research 8.

[5] Bock HH (1994) Information and entropy in cluster analysis. In Bozdogan H. et al., Kluwer Academic Press, Dordrecht

[6] Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: ACM SIGKDD, pp 269-274

[7] Ding C, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix t-factorizations for clustering. In: ACM SIGKDD, pp 126-135

[8] Govaert G, Nadif M (2003) Clustering with block mixture models. Pattern Recognition 36:463–473

[9] Govaert G, Nadif M (2005) An EM algorithm for the block mixture model. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(4):643–647

[10] Govaert G, Nadif M (2008) Block clustering with Bernoulli mixture models: Comparison of different approaches. Computational Statistics and Data Analysis 52(6):3233–3245

[11] Govaert G, Nadif M (2010) Latent block model for contingency table. Communications in Statistics - Theory and Methods 39(3):416–425

[12] Govaert G, Nadif M (2013) Co-Clustering. John Wiley & Sons

[13] Govaert G, Nadif M (2016) Mutual information, phi-squared and model-based co-clustering for contingency tables. Advances in Data Analysis and Classfication pp 1-34

[14] Hanczar B, Nadif M (2011) Using the bagging approach for biclustering of gene expression data. Neurocomputing 74(10):1595–1605

[15] Hanczar B, Nadif M (2012) Ensemble methods for biclustering tasks. Pattern Recognition 45(11):3938–3949

[16] Hartigan JA (1975) Clustering Algorithms, 99th edn. John Wiley & Sons, Inc., New York, NY, USA

[17] Hofmann T, Puzicha J (1999) Latent class models for collaborative filtering. In: IJCAI, vol 99, pp 688-693

[18] Labiod L, Nadif M (2011) Co-clustering for binary and categorical data with maximum modularity. In: ICDM, pp 1140-1145

[19] Madeira, SC, Oliveira, AL (2004). Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 1(1), 24-45.