Abstract MLSB 2017

Integrative concurrent analysis of multiple biological datasets by HALS-based multi-relational NMF

Oliver Müller-Stricker[1], Lars Kaderali[1]

[1]Institute of Bioinformatics, University Medicine Greifswald

Non-negative matrix factorization (NMF) has already been proven useful for the analysis of biological and epidemiological datasets, as e.g. gene expression, RNAi or GWAS data. For example, Kim and Tidor [1] have applied NMF to cluster genes and to predict functional relationships in yeast, Brunet et al. [2] utilized NMF to reduce the dimensionality of expression data from thousands of genes to a handful of metagenes. Hutchins et al. [3] described a novel approach to the characterization of putative regulatory sequence motifs based on NMF.

However, most approaches utilizing NMF focus on a single dataset. Yet, oftentimes it is desirable to analyze multiple interrelated datasets concurrently instead of analyzing each one separately. There are only few approaches to this task proposed in literature. For example, Wang et al. [4] introduced a technique for the prediction of protein–protein interactions from multimodal biological data sources. Gligorijević et al. [5] presented a methodology for discovery of driver genes by a holistic analysis of patient SNP data, demographic data and gene-gene interaction data.

These applications of NMF utilize an approach presented already in 2008 by Wang et al. [6] which consists of the inclusion of multiple relation matrices into the objective function of the NMF problem, the usage of matrix tri-factorization by introduction of an additional matrix which relates the two factor matrices to each other, and the addition of a graph laplacian in order to be able to include intra-type data, e.g. protein-protein interaction data.

While the methods described above are shown to yield superior results compared to other methods in the field, all of these algorithms rely on standard multiplicative updates, which oftentimes show slow convergence [7, 8]. Here, we propose a novel method for the fast integrative and concurrent analysis of interrelated datasets. For this purpose we adapt and extend the well-known Hierarchical Alternating Least Squares (HALS) algorithm for NMF [9, 10].

Our contribution is threefold. First, to be able to represent the contributions of single factor combinations, we adapt HALS for matrix tri-factorization. As stated above, an additional matrix has to be inserted in the formulation.

Second, we extend HALS to handle the factorization of multiple relations concurrently. This way it is possible to integrate data from multiple sources as well as prior knowledge directly into one single analysis.

Third, we introduce the option of weighted input into our algorithm to enable the analysis of incomplete input data, which is a common case in biological data. Here, instead of imputing unobserved input values or representing them as zero, a weight mask is introduced into the methodology.

We show the behavior of our proposed method by applying it to multi-relational biological data, as e.g. for the stratification of cancer subtypes by patient mutation profiles and expression data of ovarian, uterine and lung cancer cohorts from The Cancer Genome Atlas as well as additional molecular network data, and comparing it to current state-of-the-art NMF analysis algorithms.

References

[1] P. M. Kim and B. Tidor, “Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data,” *Genome Res.*, vol. 13, no. 7, pp. 1706–1718, Jul. 2003.

[2] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 12, pp. 4164–4169, Mar. 2004.

[3] L. N. Hutchins, S. M. Murphy, P. Singh, and J. H. Graber, “Position-dependent motif characterization using non-negative matrix factorization,” *Bioinformatics*, vol. 24, no. 23, pp. 2684–2690, Dec. 2008.

[4] H. Wang, H. Huang, C. Ding, and F. Nie, “Predicting Protein–Protein Interactions from Multimodal Biological Data Sources via Nonnegative Matrix Tri-Factorization,” *J. Comput. Biol.*, vol. 20, no. 4, pp. 344–358, Mar. 2013.

[5] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, “Patient-specific data fusion for cancer stratification and personalised treatment,” in *Biocomputing 2016*, WORLD SCIENTIFIC, 2015, pp. 321–332.

[6] F. Wang, T. Li, and C. Zhang, “Semi-Supervised Clustering via Matrix Factorization,” in *Proceedings of the 2008 SIAM International Conference on Data Mining*, 0 vols., Society for Industrial and Applied Mathematics, 2008, pp. 1–12.

[7] H. Kim and H. Park, “Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method,” *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, Jan. 2008.

[8] J. Kim and H. Park, “Fast Nonnegative Matrix Factorization: An Active-Set-Like Method and Comparisons,” *SIAM J. Sci. Comput.*, vol. 33, no. 6, pp. 3261–3281, Jan. 2011.

[9] A. Cichocki and A.-H. Phan, “Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations,” *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E92–A, no. 3, pp. 708–721, Mar. 2009.

[10] A. Cichocki, R. Zdunek, and S. Amari, “Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization,” in *Independent Component Analysis and Signal Separation*, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. Springer Berlin Heidelberg, 2007, pp. 169–176.