Information Bottleneck for Non Co-Occurence Data Supplementary Material

Yevgeny Seldin[†] **Noam Slonim**[*] **Naftali Tishby**[†‡]

[†]School of Computer Science and Engineering
[‡]Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
[*]The Lewis-Sigler Institute for Integrative Genomics
Princeton University
{seldin,tishby}@cs.huji.ac.il, nslonim@princeton.edu

Abstract

This technical report provides supplementary material for the paper "Information Bottleneck for Non Co-Occurrence Data" [1]. The report includes run time analysis for the algorithm proposed in the "Theory" section of the paper; detailed results of the algorithm application to the ESR dataset reported in the "Applications" section of the paper; visualization of the solution of MovieLens data clustering, also reported in the "Applications" section of the paper; and an additional application of the algorithm to a small dataset only mentioned in the paper.

1 Supplementary material for the Theory section

This section includes run time analysis and computational optimization proposal for the data analysis algorithm described in the "Information Bottleneck for Non Co-Occurrence Data" paper [1]. For the sake of convenience we provide the algorithm once again:

1. Start with a random (hard) partition $q(c|x)$, $q(d|y)$.
2. Iteratively until convergence (no changes at step (b) are done) traverse all rows x and columns y of a matrix in a random order. For each row/column:
   (a) Draw x (or y) from its cluster.
   (b) Reassign it to a new cluster c^* (or d^*), so that $\mathcal{L}_{min}$ is minimized. The new cluster may appear to be the old cluster, and then no change is counted.

1.1 Run-time analysis

We now analyze the run time of the algorithm. The loop (2) is executed until convergence, which in the experiments usually takes 10-40 iterations. Inside the loop we make a single pass over all $n+m$ rows and columns of a matrix. For each row x (or column y) we test the value of $\mathcal{L}_{min}$ for every possible assignment of x (or y) to a cluster c (or d). Evaluation of $\mathcal{L}_{min}$ requires evaluation of the mutual informations $I(X;C)$, $I(Y;D)$, and $I(C,D;Z)$. Evaluation of $I(X;C)$ and $I(Y;D)$ is computationally simple (for the hard assignments $I(X;C) = H(C)$ and requires only $O(|C|)$ operations). Evaluation of $I(C,D;Z)$ is more demanding, but one can evaluate $I(C,D;Z)$ for all possible assignments of a given x to all clusters c with only two computations of $I(C,D;Z)$ by using computational optimization described below. Single evaluation of $I(C,D;Z)$ requires $O(|C||D||Z|)$ operations, where $|Z|$ is the cardinality of Z (5 in most applications considered). Thus in total we obtain a complexity of $O((n+m)|C||D|)$ for a single pass through loop (2).

1.2 Computational Optimization

A strait-forward calculation of $I(C,D;Z)$ is relatively demanding. We consider the following derivation, which is similar to the optimization trick used in [2, 3]:

$$\arg \min_{q(c|x),q(d|y)} I(X;C) + I(Y;D) - \beta I(C,D;Z)$$

$$= \arg \min_{q(c|x),q(d|y)} I(X;C) + I(Y;D) - \beta(H(Z) - H(Z|C,D))$$

$$= \arg \min_{q(c|x),q(d|y)} I(X;C) + I(Y;D) + \beta H(Z|C,D)$$

since $H(Z)$ is constant. Now:

$$H(Z|C,D) = \sum_c \hat{q}(c) H(Z|c,D)$$

We denote by H_{old} the entropies after a sample x was taken out of its cluster, by $q_{old}(c)$ the corresponding distribution over c, and by H_{c_x} the entropies after x was assigned to a trial cluster c_x and q_{c_x} the corresponding distribution over c. Then one obtains:

$$H_{old}(Z|C,D) = \sum_c \hat{q}_{old}(c) H_{old}(Z|c,D)$$

$$H_{c_x}(Z|C,D) = \sum_c \hat{q}_{c_x}(c) H_{c_x}(Z|c,D)$$

$$= \hat{q}_{c_x}(c_x) H_{c_x}(Z|c_x,D) + \sum_{c \neq c_x} \hat{q}_{c_x}(c) H_{c_x}(Z|c,D)$$

$$= \hat{q}_{c_x}(c_x) H_{c_x}(Z|c_x,D) + \sum_{c \neq c_x} \hat{q}_{c_x}(c) H_{old}(Z|c,D) = (*)$$

Since clusters except c_x passed no changes. We now observe that for clusters other than c_x:

$$\hat{q}_{c_x}(c) = (1 - \hat{p}(x)) \hat{q}_{old}(c)$$

Thus:

$$(*) = \hat{q}_{c_x}(c_x) H_{c_x}(Z|c_x,D) + (1 - \hat{p}(x)) \sum_{c \neq c_x} \hat{q}_{old}(c_x) H_{old}(Z|c,D)$$

$$= \hat{q}_{c_x}(c_x) H_{c_x}(Z|c_x,D) + (1 - \hat{p}(x))(H_{old}(Z|C,D) - \hat{q}_{old}(c_x) H_{old}(Z|c_x,D))$$

Which means that in order to evaluate all possible assignments of x to clusters c we need to compute the entropy $H(Z|C,D)$ only twice, and a similar analysis applies to reassignment of columns y to column clusters d.

2 Supplementary material for the Applications section

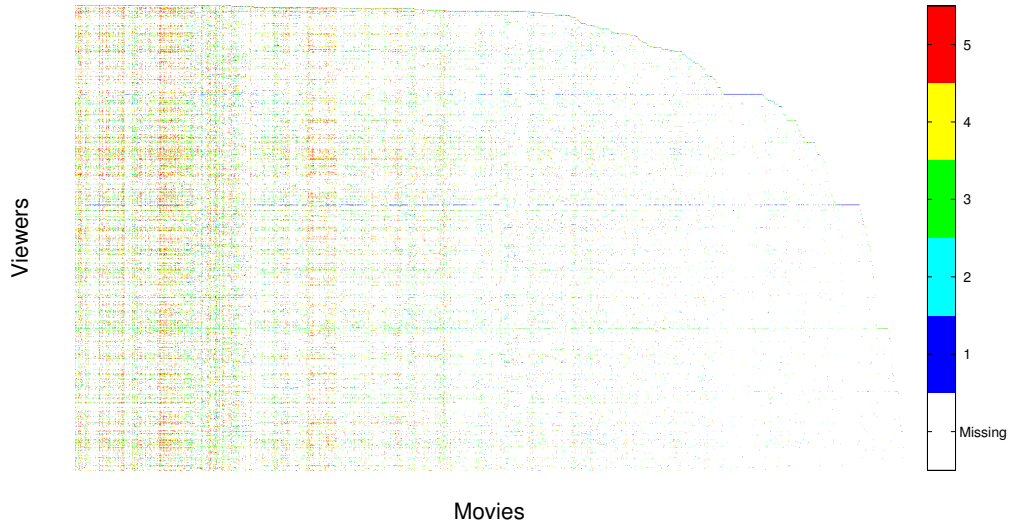
2.1 One Dimensional Clustering - Comparison to I-Clust

In this section we provide detailed coherence results of the solutions obtained for ESR dataset as reported in the Applications section of the "Information Bottleneck for Non Co-Occurrence Data" paper. The detailed results include evaluation of the coherence according to the three Gene Ontology annotations [5] and their average - see Table 1. (In the paper only the average is reported.)

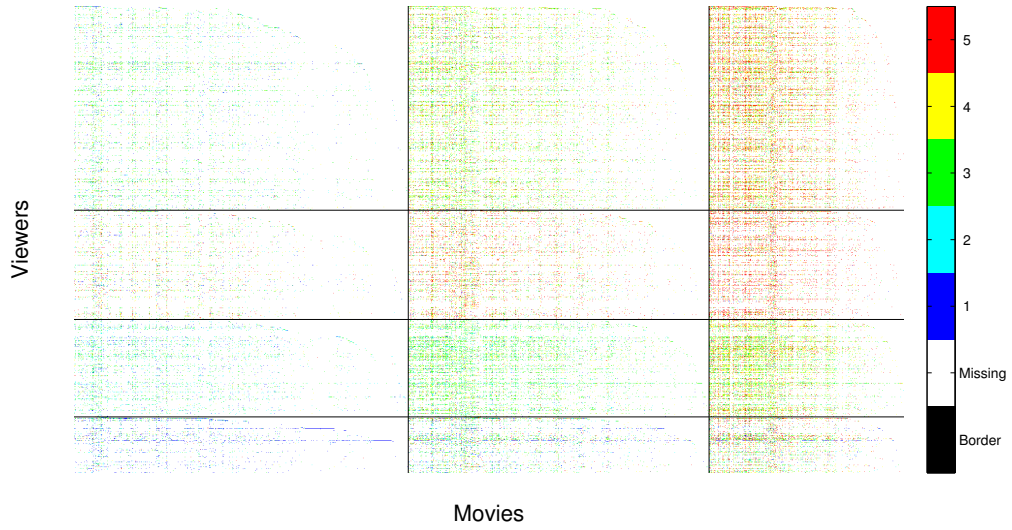
2.2 Matrix Completion and Collaborative Filtering

This subsection provides an illustration of the MovieLens rating data[1] and its clustering into $|C| = 4$ clusters of viewers and $|D| = 3$ clusters of movies obtained by the algorithm suggested in the paper (see Figure 1). It may be observed that the original mixed data is clustered into patches of roughly homogeneous ratings.

[1] http://www.grouplens.org



(a) Original data.



(b) Data resorted according to obtained clustering.

Figure 1: MovieLens data before (a) and after (b) clustering. Matrix rows correspond to movies, columns to viewers and ratings are depicted with colors when present and white when absent (see colormaps on the right). The figure illustrates the original data (a) and the same data after permutation of rows and columns with accordance to clustering into $|C| = 4$ clusters of viewers and $|D| = 3$ clusters of movies (b). It may be observed that the original data is clustered into patches with roughly homogeneous ratings.

Table 1: **Detailed comparison of clusters coherence for the ESR dataset according to the tree GO annotations.** The table provides detailed coherence results for the achieved solutions for $N_c = 5, 10, 15$ and 20 row clusters according to the three GO annotations. The results achieved by the Iclust algorithm [4] are shown in brackets next to the results of our algorithm.

Annotation	$N_c = 5$	$N_c = 10$	$N_c = 15$	$N_c = 20$
BP	65 (75)	54 (50)	52 (51)	53 (51)
MF	69 (77)	47 (43)	52 (54)	39 (41)
CC	73 (86)	57 (53)	45 (52)	35 (33)
Average	69 (79)	53 (49)	50 (52)	42 (42)

2.3 Small Datasets

To demonstrate the ability of our algorithm to cope with datasets with a small number of columns, we use the well known Iris dataset [6] which consists of four numeric parameters for 150 Iris flowers belonging to three classes of size 50 each. We quantize the values in each column into three equally populated bins. The quantized matrix is clustered into three classes and only 7 misclassifications are obtained. This example demonstrates a principled advantage of our algorithm over I-Clust [4] that requires availability of a much larger amount of matrix columns to be applied.

References

[1] Yevgeny Seldin, Noam Slonim, and Naftali Tishby. Information bottleneck for non co-occurrence data. In B. Schölkopf, J.C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.

[2] Noam Slonim, Nir Friedman, and Naftali Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the* 25^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2002.

[3] Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate information bottleneck. *Neural Computation*, 18, 2006.

[4] Noam Slonim, Gurinder Singh Atwal, Gasper Tracik, and William Bialek. Information-based clustering. In *Proceedings of the National Academy of Science (PNAS)*, volume 102, pages 18297–1830, Dec. 2005.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, May 2000.

[6] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II:179–188, 1936.