Local Topic Discovery via Boosted Ensemble of Nonnegative Matrix Factorization

Sangho Suh
Korea University
Seoul, South Korea
sh31659@gmail.com

Jaegul Choo
Korea University
Seoul, South Korea
jchoo@korea.ac.kr

Joonseok Lee
Google Research
Mountain View, CA, USA
joonseok@google.com

Chandan K. Reddy
Virginia Tech
Arlington, VA, USA
reddy@cs.vt.edu

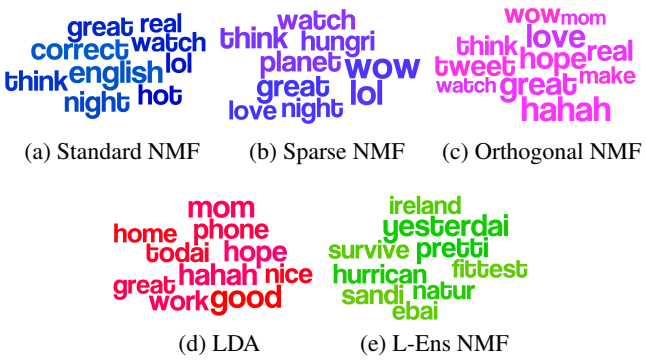
Abstract

Nonnegative matrix factorization (NMF) has been increasingly popular for topic modeling of large-scale documents. However, the resulting topics often represent only general, thus redundant information about the data rather than minor, but potentially meaningful information to users. To tackle this problem, we propose a novel ensemble model of nonnegative matrix factorization for discovering high-quality local topics. Our method leverages the idea of an ensemble model to successively perform NMF given a residual matrix obtained from previous stages and generates a sequence of topic sets. The novelty of our method lies in the fact that it utilizes the residual matrix inspired by a state-of-the-art gradient boosting model and applies a sophisticated local weighting scheme on the given matrix to enhance the locality of topics, which in turn delivers high-quality, focused topics of interest to users.[1]

1 Introduction

Until recently, the main focus of the two major topic modeling approaches—i.e., probabilistic and matrix factorization methods—has been to find a given number of bases or probability distributions, which we call topics, over the dictionary so that they can explain a given set of documents. Most of the existing topic modeling methods focus on generating global topics to explain the majority of a given document corpus. However, such topics do not often provide meaningful information and are sometimes even redundant with each other when multiple similar topics are dominant in the corpus.

For instance, Fig. 1 shows the sample topics generated by existing algorithms and our proposed method called L-EnsNMF, using Twitter dataset collected from New York City. Clearly, the keywords found by the baseline methods are shown to be general but not informative–see words, such as 'lol,' 'wow,' 'great,' and 'hahah.' On the contrary, our method, which attempts to discover local topics returns specific and insightful keywords, e.g., 'ireland,' 'survive,' and



(a) Standard NMF (b) Sparse NMF (c) Orthogonal NMF

(d) LDA (e) L-Ens NMF

Figure 1: Topics computed from Twitter dataset

'hurricane sandi'–which devastated New York City in 2012. Searching for 'ireland hurricane sandy' on the web, we discovered the local news that the Ireland football team visited New York in June 2013 to support a community hit by Hurricane Sandy. Moreover, 'hurricane sandi' were not found in any of the 100 topics (ten keywords each) generated by any existing methods, showing that a local topic discovery approach does not only enhance topic quality but also allows the discovery of potentially meaningful topics that would otherwise be left undiscovered.

In response, this paper proposes a local ensemble model of nonnegative matrix factorization (NMF) [Lee and Seung, 1999]. Although NMF has been used previously in the ensemble framework in machine learning applications, including clustering [Greene *et al.*, 2008], classification [Wu *et al.*, 2015], and bioinformatics [Yang *et al.*, 2014], most of these existing ensemble methods primarily focus on aggregating the outputs from multiple individual models constructed independently with some variations on input matrices and other parameter settings. Thus, they are not applicable in topic modeling where we focus on the learned bases themselves. Furthermore, none of them have tackled the idea of constructing an ensemble of NMF models based on a gradient boosting framework, which grants a clear novelty of our work. TF-IDF is also known to work reasonably well for what our approach aims to achieve, i.e., down-weighting general, redundant words and up-weighting infrequent, specific words in a document corpus. However, TF-IDF can distort the semantic

[1]This paper is an abridged version of an already published paper [Suh *et al.*, 2016], invited to IJCAI'17 Sister Conferences track for a broader audience.

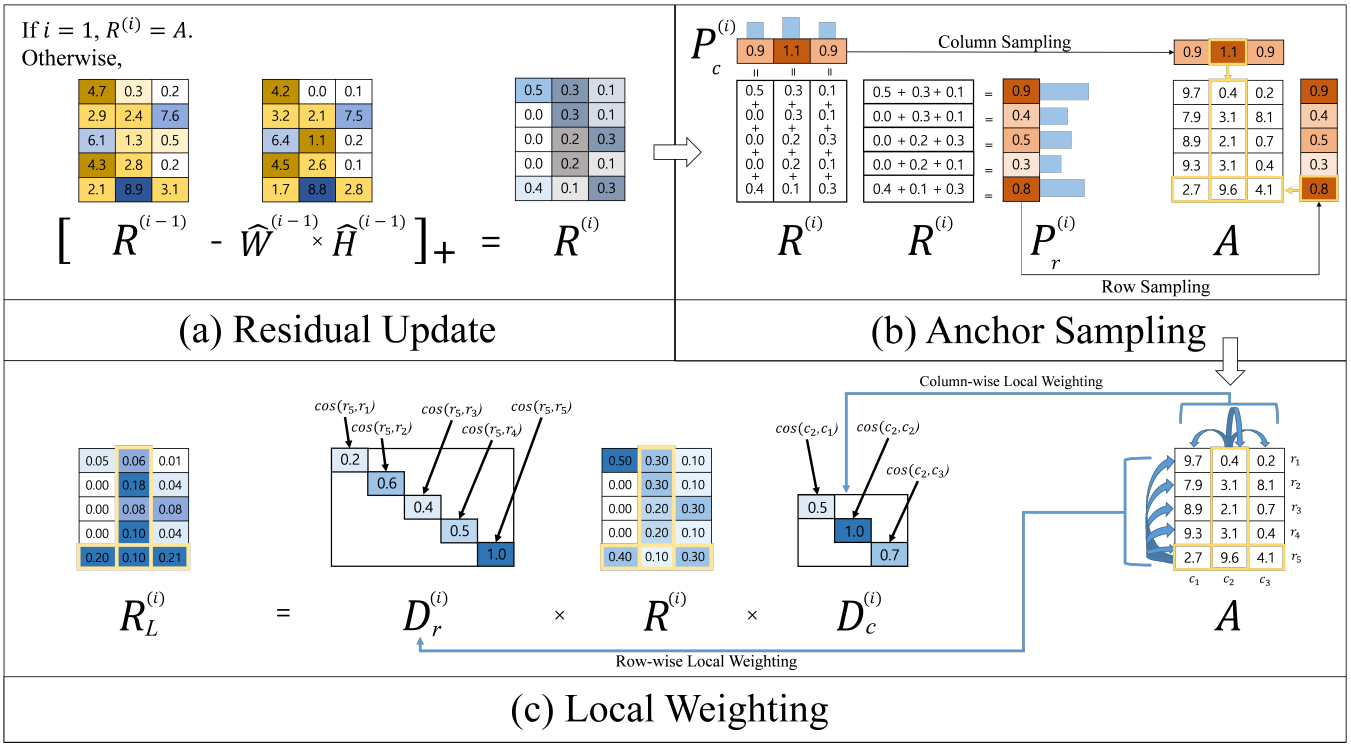


Figure 2: Overview of our local ensemble NMF (L-EnsNMF) approach

coherence of each topic, and thus it is not used much in the context of topic modeling.

Overall, the main contributions of this paper are as follows:

- A novel ensemble of nonnegative matrix factorization method based on a gradient boosting framework.
- An extensive quantitative analysis with various document datasets, showing the superiority of our method.

2 NMF for Topic Modeling

Given a nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$, and an integer $k \ll \min(m, n)$, nonnegative matrix factorization (NMF) [Lee and Seung, 1999] finds a lower-rank approximation given by

$$X \approx WH, \tag{1}$$

where $W \in \mathbb{R}_+^{m \times k}$ and $H \in \mathbb{R}_+^{k \times n}$ are nonnegative factors. NMF is typically formulated in terms of the Frobenius norm as

$$\min_{W,\, H \geq 0} \|X - WH\|_F^2 . \tag{2}$$

where '$\geq$' applies to every element of the given matrix in the left-hand side. In topic modeling, $x_i \in \mathbb{R}_+^{m \times 1}$, the i-th column of X, corresponds to the bag-of-words representation of document i with respect to m keywords. k denotes the number of topics. $w_l \in \mathbb{R}_+^{m \times 1}$, the l-th nonnegative column vector of W, represents the l-th topic as a weighted combination of m keywords. The i-th column vector of H, $h_i \in \mathbb{R}_+^{k \times 1}$, represents document i as a weighted combination of k topics.

3 L-EnsNMF for Local Topic Modeling

We propose our gradient-boosted local ensemble NMF approach called L-EnsNMF.[2] As shown in Fig. 2, L-EnsNMF iteratively performs three steps, residual update, anchor sampling, and local weighting. In this ensemble model, an individual learner corresponds to NMF. That is, given a nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$, we learn an additive model $\hat{X}^{(q)}$ with q products $W^{(i)}H^{(i)}$:

$$X \approx \hat{X}^{(q)} = \sum_{i=1}^{q} W^{(i)} H^{(i)} \tag{3}$$

where $W^{(i)} \in \mathbb{R}_+^{m \times k_s}$, $H^{(i)} \in \mathbb{R}_+^{k_s \times n}$ and q is the number of individual learners. In other words, the i-th *stage* represents a local NMF model discovering the i-th set of k_s local topics. To achieve this approximation, we use the Frobenius norm-based objective function as follows:

$$\min_{W^{(i)}, H^{(i)} \geq 0,\, i=1,\cdots,q} \left\| X - \sum_{i=1}^{q} W^{(i)} H^{(i)} \right\|_F^2 . \tag{4}$$

Residual Update. L-EnsNMF solves this problem in a forward stage-wise manner [Hastie *et al.*, 2009], which iteratively adds a new local model to better approximate X, fitting the i-th local NMF, $W^{(i)}H^{(i)}$, with rank k_s to the localized residual, i.e., the unexplained portion from previously learned

[2] The code is available at `https://github.com/sanghosuh/lens_nmf-matlab`

$i-1$ local models. To this end, we define the (non-localized) nonnegative residual matrix at stage i as

$$R^{(i)} = \begin{cases} X & \text{if} \quad i = 1 \\ \left[R^{(i-1)} - W^{(i-1)}H^{(i-1)}\right]_+ & \text{if} \quad i \geq 2 \end{cases} \quad (5)$$

where $[\cdot]_+$ is an operator that converts every negative element in the matrix to zero.

Anchor Sampling. Next, we use $P_r^{(i)}$ and $P_c^{(i)}$ to assign higher weights to those rows or columns less explained (large residuals) by previous stages. Let us define the probability distributions $P_r^{(i)}$ and $P_c^{(i)}$ over row indices, x's, and over column indices, y's, respectively, as

$$P_r^{(i)}(x) = \frac{\sum_{s=1}^{n} R^{(i)}(x,s)}{\sum_{l=1}^{m}\sum_{s=1}^{n} R^{(i)}(l,s)} \text{ for } x = 1, \cdots, m \quad (6)$$

$$P_c^{(i)}(y) = \frac{\sum_{l=1}^{m} R^{(i)}(l,y)}{\sum_{l=1}^{m}\sum_{s=1}^{n} R^{(i)}(l,s)} \text{ for } y = 1, \cdots, n. \quad (7)$$

In the above equations, higher probability values are assigned to those rows or columns with larger values in residual matrix $R^{(i)}$. That is, a higher probability indicates that the corresponding row or column is less explained up to the previous stage. As illustrated in Fig. 2(b), we sample from this probability distribution a single row a_r and a column a_c, which we call an *anchor point*, which corresponds to a particular keyword and a document that were not yet sufficiently explained in previous stages, respectively.

Local Weighting. We then apply local weighting on the residual matrix $R^{(i)}$ to obtain its localized version $R_L^{(i)}$. Given a local residual matrix $R_L^{(i)}$ at stage i, we obtain the term-by-topic matrix $\hat{W}^{(i)}$ and the topic-by-document matrix $\hat{H}^{(i)}$ by solving

$$\left(W^{(i)}, H^{(i)}\right) = \arg\min_{W,H\geq 0} \left\| R_L^{(i)} - WH \right\|_F^2. \quad (8)$$

Here, the localized residual matrix $R_L^{(i)}$ is formed as

$$R_L^{(i)} = D_r^{(i)} R^{(i)} D_c^{(i)}, \quad (9)$$

where diagonal matrices $D_r^{(i)} \in \mathbb{R}_+^{m\times m}$ and $D_c^{(i)} \in \mathbb{R}_+^{n\times n}$ perform row- and column-wise scaling, respectively, as shown in Fig. 2(c).

The diagonal entries of $D_r^{(i)}$ and $D_c^{(i)}$ are computed based on the similarity of each row and column to the anchor row a_r and column a_c, respectively. Specifically, given the selected a_r and a_c, we use the cosine similarity to compute the l-th diagonal entry of $D_r^{(i)}(l,l)$ and the s-th diagonal entry of $D_c^{(i)}(s,s)$, respectively, as

$$D_r^{(i)}(l,l) = \cos\left(X(a_r,:), X(l,:)\right) \text{ for } l = 1, \cdots, m \quad (10)$$

$$D_c^{(i)}(s,s) = \cos\left(X(:,a_c), X(:,s)\right) \text{ for } s = 1, \cdots, n. \quad (11)$$

After formulating the localized residual matrix as described above, we use $R_L^{(i)}$ (Eq. (9)) as an input matrix for NMF and obtain $W^{(i)}$ and $H^{(i)}$, as in Eq. (8). When computing the residual matrix in the next stage using $W^{(i)}$ and $H^{(i)}$, as shown in Eq. (5), however, it may end up removing only the fraction of the residuals, which can be significantly smaller than the unweighted residuals since all the weights are less than or equal to 1. To adjust this shrinking effect caused by local weighting, we recompute $H^{(i)}$ using the given $W^{(i)}$ and the non-weighted residual matrix $R^{(i)}$, i.e.,

$$H^{(i)} = \arg\min_{H\geq 0} \left\| W^{(i)}H - R^{(i)} \right\|_F^2. \quad (12)$$

In this manner, our approach still maintains the localized topics $W^{(i)}$ from $R_L^{(i)}$ while properly subtracting the full portions explained by these topics from $R^{(i)}$ for the next stage.

Fast Rank-2. In addition to the novel local weighting scheme, we apply a recently proposed, highly efficient NMF algorithm based on an active-set method [Kuang and Park, 2013] for a small rank value.

Table 1: Summary of the data sets used.

	Reuters	20News	Enron	VisPub	Twitter
#docs	7,984	18,221	2,000	2,592	2,000
#words	12,411	36,568	19,589	7,535	4,212

4 Experiments

We selected the following five real-world document datasets: 1) Reuters-21578 (**Reuters**),[3] 2) 20 Newsgroups (**20News**),[4] 3) **Enron**,[5] 4) IEEE-Vis (**VisPub**),[6] and 5) **Twitter**, as summarized in Table 1.

Using these datasets, we compared our L-EnsNMF against various state-of-the-art methods, including standard NMF (**StdNMF**) [Kim *et al.*, 2014],[7] sparse NMF (**SprsNMF**) [Kim and Park, 2007],[8] orthogonal NMF (**OrthNMF**) [Ding *et al.*, 2006],[9] and latent Dirichlet allocation (**LDA**) [Blei *et al.*, 2003].[10] In most of these methods, we used default parameter values provided by the corresponding software library.

All the experiments were conducted using MATLAB 8.5 (R2015a) on a desktop computer with dual Intel Xeon E5-2687W processors.

4.1 Quantitative Analysis

Total Document Coverage. This measure computes how many documents (out of the entire document set) can be ex-

[3] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[4] http://qwone.com/~jason/20Newsgroups/

[5] https://www.cs.cmu.edu/~./enron/

[6] http://www.vispubdata.org/site/vispubdata/

[7] https://github.com/kimjingu/nonnegfac-matlab

[8] http://www.cc.gatech.edu/~hpark/software/nmf_bpas.zip

[9] http://davian.korea.ac.kr/myfiles/list/Codes/orthonmf.zip

[10] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

Table 2: Total document coverage results on VisPub dataset. The reported results are averaged values over 20 runs. The best performance values are shown in **bold**, and the second best ones are underlined.

	$k = 50$ ($k_s = 2$, $q = 25$)				
Number of keywords	Std NMF	Sprs NMF	Orth NMF	LDA	L-Ens NMF
3	0.962	0.951	0.963	**0.977**	0.972
4	0.770	0.717	0.772	**0.902**	0.892
5	0.428	0.367	0.435	0.651	**0.689**
6	0.155	0.125	0.158	0.336	**0.412**
7	0.039	0.030	0.040	0.107	**0.178**
8	0.007	0.006	0.007	0.028	**0.057**
9	0.001	0.001	0.001	0.001	**0.012**
10	0.000	0.000	0.000	0.000	**0.003**
Average	0.295	0.275	0.297	0.375	**0.402**
	$k = 100$ ($k_s = 2$, $q = 50$)				
Number of keywords	Std NMF	Sprs NMF	Orth NMF	LDA	L-Ens NMF
3	0.962	0.948	0.962	0.979	**0.980**
4	0.724	0.676	0.722	**0.919**	0.889
5	0.346	0.303	0.345	**0.676**	0.669
6	0.111	0.099	0.111	0.336	**0.397**
7	0.028	0.024	0.028	0.105	**0.179**
8	0.007	0.005	0.007	0.024	**0.060**
9	0.002	0.001	0.001	0.003	**0.017**
10	0.000	0.000	0.000	0.000	**0.005**
Average	0.273	0.257	0.272	0.380	**0.400**

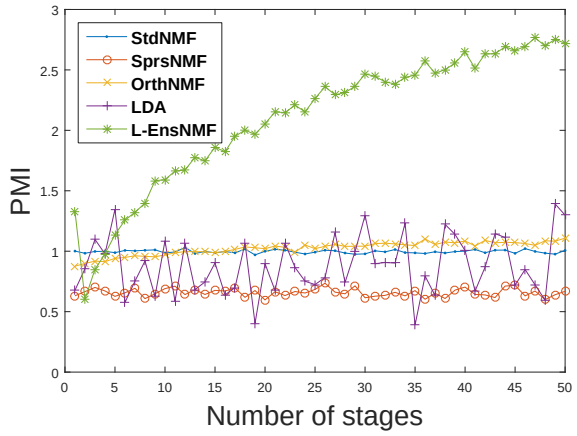


Figure 3: Topic coherence values across various stages when 100 topics ($k_s = 2$, $q = 50$) are computed for VisPub dataset. Each value of our method represents the average topic coherence value of k_s corresponding topics per stage. The results of the other methods show the average values per k_s topics. The results were obtained by computing the average values over 1,000 runs.

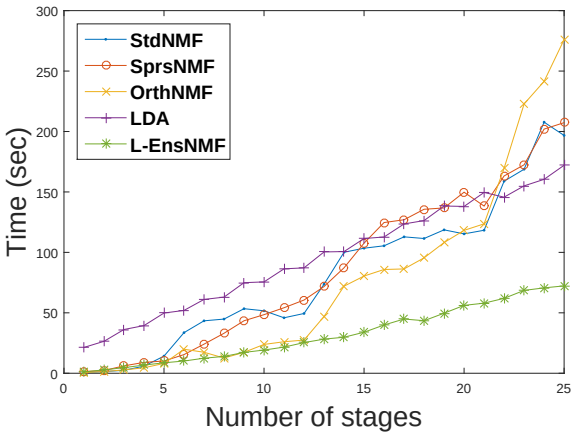


Figure 4: Comparison of computing times for VisPub dataset. The results were obtained from the average values over 50 runs.

plained by a given set of topics. Here, a document is said to be "*explained*" if there exists a topic such that at least a certain number of keywords among its most representative keywords are found in that document. In Table 2, our method is shown to be the best or the second best method for all the different numbers of topics.

Topic Coherence. Fig. 3 shows how the topic coherence values change as the stages proceed in our ensemble model. One can see that the topic coherence is constantly improved as the stages proceed, which generates topics with much better quality than other methods. This supports our claim that the gradient boosting-based ensemble framework for NMF works well in topic modeling applications and that the topics generated in later stages in this framework have significant advantages than those generated by other existing methods.

Computing Times. We measured the running time of different methods by changing the total number of topics, k, from 2 to 50. In the case of our ensemble NMF method, we fixed k_s as 2 while changing q from 1 to 25. As shown in Fig. 4, our method runs fastest, and more importantly, it scales better than other methods with respect to k since its computational complexity, i.e., k_s, stays constant at 2 regardless of the total number of topics.

5 Future Work

We plan to expand our work to an interactive topic discovery system [Choo *et al.*, 2013; Kim *et al.*, 2017] by flexibly steering the local weighting process in a user-driven manner so that the subsequent topics can properly reflect a user's subjective interest and task goals.

Acknowledgements

This work was supported in part by the National Science Foundation grants IIS-1707498, IIS-1619028, IIS-1646881 and by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2016R1C1B2015924). Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of funding agencies.

References

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

[Choo *et al.*, 2013] Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):1992–2001, 2013.

[Ding *et al.*, 2006] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[Greene *et al.*, 2008] Derek Greene, Gerard Cagney, Nevan Krogan, and Pádraig Cunningham. Ensemble non-negative matrix factorization methods for clustering protein–protein interactions. *Bioinformatics*, 24(15):1722–1728, 2008.

[Hastie *et al.*, 2009] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[Kim and Park, 2007] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[Kim *et al.*, 2014] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[Kim *et al.*, 2017] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2017.

[Kuang and Park, 2013] Da Kuang and Haesun Park. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 739–747, 2013.

[Lee and Seung, 1999] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Suh *et al.*, 2016] Sangho Suh, Jaegul Choo, Joonseok Lee, and Chandan K Reddy. L-EnsNMF: Boosted local topic discovery via ensemble of nonnegative matrix factorization. 2016.

[Wu *et al.*, 2015] Qingyao Wu, Mingkui Tan, Xutao Li, Huaqing Min, and Ning Sun. Nmfe-sscc: Non-negative matrix factorization ensemble for semi-supervised collective classification. *Knowledge-Based Systems*, 89:160–172, 2015.

[Yang *et al.*, 2014] Peng Yang, Xiaoquan Su, Le Ou-Yang, Hon-Nian Chua, Xiao-Li Li, and Kang Ning. Microbial community pattern detection in human body habitats via ensemble clustering framework. *BMC systems biology*, 8(Suppl 4):S7, 2014.