The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Ensemble Non-negative Matrix Factorization for Clustering Biomedical Documents

Shanfeng Zhu[1,2,*] Wei Yuan[1,2] Fei Wang[1,2]

[1]School of Computer Science and Technology, Fudan University, Shanghai 200433, China
[2]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China

Abstract Searching and mining biomedical literature database, such as MEDLINE, is the main source of generating scientific hypothesis for biomedical researchers. Through grouping similar documents together, clustering techniques can facilitate user's need of effectively finding interested documents. Since non-negative matrix factorization (NMF) can effectively capture the latent semantic space with non-negative factorization in both the basis and the weight, it has been utilized to clustering general text documents. Considering the stochastic nature of NMF with respect to initialization, we propose to use ensemble NMF for biomedical document clustering. The performance of ensemble NMF was evaluated on clustering a large number of datasets generated from TREC Genomics track dataset. The experimental results show that our method outperforms classical clustering algorithms bisect k-means, k-means and hierarchical clustering significantly in most of the datasets.

1 Introduction

Through indexing over 16 million biomedical documents, MEDLINE [10] has accumulated more than 40 years' scientific findings in biomedical domain, and thus becomes the main source of generating scientific hypothesis and discovering new knowledge [3]. Not surprisingly, researchers are usually overwhelmed by the large number of available literature information. Clustering techniques can alleviate this problem by grouping similar documents together. Some researchers have already carried out studies on clustering biomedical documents. Lee et al. made use of hierarchical clustering techniques (group-wise average and single pass clustering) to cluster 15,405 articles cited in OMIM database, and an additional dataset of 56 articles cited in four biological review articles[6]. Yoo and Hu compared various document clustering approaches, such as K-means, Bisecting K-means, Suffix Tree Clustering (STC) and three hierarchical methods (single-link, complete-link and average-link) on clustering collected MEDLINE documents on different diseases[12]. They found that partional clustering techniques outperform hierarchical clustering techniques significantly in the experiment.

Similar to SVD, non-negative matrix factorization(NMF) is also a kind of dimension reduction technique. However it has distinct features of preserving the structure of the original data and keeping the non-negativity in both basis and weight. It was first proposed

*zhushanfeng@gmail.com

by Lee and Sung for tackling the problems in image processing[4], and later found its application in many other domains, such as information retrieval. Xu et al. made use of NMF to clustering general text documents using the TDT2 and the Reuters document corpora[11]. They found that NMF based method surpasses the latent semantic indexing method and the spectral clustering method in the experiment. In document clustering based on NMF, the latent semantic space has a very intuitive explanation, where each axis stands for the basis topic of a particular cluster, and each document is represented by the additive combination of different basis topics. A document can be easily grouped into the cluster where it has the largest projection value.

Inspired by these studies, we first make use of NMF to clustering biomedical text documents, which has not been examined yet. Second, considering the stochastic characteristic of NMF, we propose to use ensemble clustering to get a consensus result from different trials of NMF with different initialization parameters. It makes NMF less sensitive to initialization parameters, and thus can produce robust result. Finally, a new parameter updating algorithm based on projected gradient method[7] is used to speedy the convergency. The performance of ensemble NMF has been examined on a large number of datasets. The experimental results show that the performance of ensemble NMF outperforms classical clustering algorithms bisect k-means, k-means and hierarchical clustering significantly in most of the datasets.

The rest of paper is organized as follows: In section 2, we introduce the NMF and ensemble NMF for biomedical document clustering. The testing data set and the experimental result are described in Section 3. Finally we make a conclusion and discuss the future works.

2 Ensemble Clustering based on NMF

Assume that there are n documents $D = \{d_1, d_2, \cdots, d_n\}$ and m distinct terms $W = \{w_1, w_2, \cdots, w_m\}$ (after removing stopping words and words stemming) in the corpus, we can represent D by a matrix $A_{m\times n}$ and each document d_i by a vector A_i with $tf-idf$ weighting scheme. That is $A_i = [a_{1i}, a_{2i}, \cdots, a_{mi}]^T$, and $a_{ji} = t_{ji} \times log(n/idf_j)$, where idf_j, t_{ji} stand for the number of documents containing term w_j and the frequency of w_j in document d_i. A_i is then normalized to have unit Euclidean length, and becomes the i-th column of A. We can use NMF to factorize A to get the clustering result.

2.1 NMF

Assume that D consists of k clusters, the goal of NMF is to factorize A into the product of two non-negative matrix, the base matrix $U_{m\times k}$ and the weight(coefficient) matrix $V^T_{k\times n}$, and try to minimize the Frobenius norm of the difference between $A - UV^T$. That is to minimize the following objective function: $J = \|A - UV^T\|^2_F$ with the constraints of $u_{ij} \geq 0, v_{xy} \geq 0$, where $0 \leq i \leq m$, $0 \leq j \leq k$, $0 \leq x \leq n$, and $0 \leq y \leq k$.

This is a nonlinear optimization problem, and has been proved to be a NP-hard problem [9]. The most popular heuristical algorithm is multiplicative update rule, where U and V are randomly initialized, and they are updated using expectation maximization algorithm [5].

$$u_{ij} \leftarrow u_{ij} \frac{(AV)_{ij}}{(UV^TV)_{ij}}$$

$$v_{ij} \leftarrow v_{ij} \frac{(A^T U)_{ij}}{(VU^T U)_{ij}}$$

This updating algorithm is easy to implement and usually obtains good result, but converges very slow. Recently Lin applied projected gradient methods to NMF that converges much faster than multiplicative update rule [7]. We also make use of this method in our experiment. After obtaining the coefficient matrix $V_{n\times k}$, we assign the cluster label to each document where it has the largest weight. That is, assign cluster x to d_i if $x = \arg\max_j v_{ij}$.

Xu et al found that a normalized cut weighted form of NMF outperforms ordinary NMF in the experiment[11]. In this case, we need to calculate diagonal matrix $Z = diag(A^T Ae)$, and use $A' = AZ^{-1/2}$ for further factorization. Both NMF and weighted NMF will be evaluated in our work.

2.2 Ensemble NMF

Due to its stochastic nature, the result of NMF relies on random initialization. As a result, we will get different clustering solutions with same dataset across different runs. To obtain a consensus clustering result, we make use of ensemble clustering methods to aggregate various clustering results that come from different initialization parameters in NMF. Many ensemble clustering algorithms have been proposed to tackle this problem, and here we use MCLA (Meta-CLustering Algorithms) proposed by Strehl and Ghosh for aggregating clustering results[8]. In MCLA, each cluster is represented by a hyperedge, and we try to group and combine related hyperedges. Each document is assigned to a hypeedge where it occurs most often. For the detail of MCLA, please refer to [8]

Here is the overview of Ensemble NMF for clustering biomedical document corpus D.

(1) Calculate the $tf-idf$ weighted term document matrix A with unit Euclidean distance for each row.

(2) For i=1 to τ

- Randomly initialize U and V

- Use NMF to factorize A to get U and V

- Determine the individual clustering result C_i according to V

(3) Use MCLA to obtain the consensus clustering result C from the set of individual clustering result $\{C_1, C_2, \cdots, C_\tau\}$

In this work we set τ to 50 where a stable clustering result usually can be obtained. In addition, the number of true clusters k in the dataset is given as a prior parameter to NMF.

3 Experiment Result

3.1 Experimental Dataset

We created the experimental dataset from the document corpora of TREC Genomics track 2004[2], which was composed of a set of 4,591,008 MEDLINE documents. In the ad hoc retrieval task, 50 topics were distributed to participating retrieval systems as queries. Biologists assessed the relevance of retrieved records from each participant retrieval system, and obtained a set of relevant documents for each topic with high reliability. After removing small size topics ($<$10 relevant documents) and the documents associating with

Table 1: Summary of statistical characteristics of Genomics2004

Collection	Data	N_d	W	K	N_l	Balance
Genomics2004	T200410a	1176	5460	10	395.3	0.0664
Genomics2004	min of all 80	133	1461	3	310.1	0.0216
Genomics2004	max of all 80	1757	6467	10	465.5	0.5625
Genomics2004	mean of all 80	739.1	3850.4	6.5	379.0	0.1198

more than one topic, we obtained a base dataset of 4400 documents in 39 topics. For a robust comparison, we built 80 different datasets from this base dataset by randomly selecting 3 to 10 topics. With a specific number of topics, 10 different datasets were generated. This collection of 80 datasets is referred as Geonomic2004 in this work. By keeping the most important information, we extract three informative fields, title, abstract and MeSH terms from MEDLINE to form the document. Here we extend MeSH terms from the root of MeSH tree structure to include as many information as possible. Some standard procedures have been used in the pre-processing step, such as removing stop words, case folding and stemming.

The statistical characteristics of Genomics2004 are illustrated in Table 1, where N_d is the number of documents, W is the number of distinct words(tokens), K is the number of topics(classes), N_l is the average number of words in each document, and balance is the size ratio of the smallest class to the largest class. We name each dataset in the collection by combining an initial alphabet "T", the year, the number of topics, and the order of the dataset. For example, "T2004010a" represents the first dataset with 10 topics generated from the Genomics track 2004. We can see that Genomics2004 collection varies significantly in some important characteristics: the number of documents in each dataset varies from 133 to 1757, the number of words from 1461 to 6467, the average length of document from 310.1 to 465.5 and the balance from 0.0216 to 0.5625. This great variety makes it highly suitable for comparing the performance of different clustering algorithms.

3.2 Evaluation Metric

In this work, since we have true label of each document, external measures can be used as evaluation criteria. Among several well-known external measures, mutual information is found as a superior measure over purity, average entropy and F-measure[1]. NMI is computed according to the following formula,

$$\text{NMI} = \frac{I(X;Y)}{\sqrt{H(X)\cdot H(Y)}},$$

where X and Y are the predicted clusters and the correct(true) class labels, respectively, $I(X;Y)$ is the mutual information between X and Y, and $H(X)$ and $H(Y)$ are the entropy of X and Y, respectively. Furthermore, Zhong and Ghosh[13, 14] proposed a sample

estimate to compute the NMI,

$$\mathrm{NMI} = \frac{\sum_{h,l} n_{h,l} \log(\frac{n \cdot n_{h,l}}{n_h n_l})}{\sqrt{(\sum_h n_h \log \frac{n_h}{n})(\sum_l n_l \log \frac{n_l}{n})}}, \tag{1}$$

where n is the total number of documents in the whole collection, n_h is the number of documents in class h (standard), n_l is the number of documents in cluster l (predicted), and $n_{h,l}$ is the number of documents in both class h and cluster l. The range of NMI is between zero and one, where an NMI value of zero means that the result is actually a random partitioning, and an NMI value close to one means that the almost perfect partitioning result is achieved.

3.3 Experimental Procedure and Result

To illustrate the effectiveness of NMF and ensemble NMF, we also obtained the clustering result using K-means, Bisect K-means and Hierarchical clustering (average-link) by CLUTO[1]. For achieving ideal clustering results for these three algorithms, the cosine similarity is used to measure the similarity between two documents. Additionally, to demonstrate the stochastic nature of NMF, we use Min-NCW-NMF to stand for the worst case of Normalized Cut Weighted NMF (NCW-NMF) out of all 50 runs, where it obtains the lowest NMI. Altogether we need to compare seven different algorithms: Hierarchal Clustering, K-means, bisect K-means, NMF, NCW-NMF, Min-NCW-NMF and Ensemble Normalized Cut Weighted NMF (EN-NCW-NMF). For each dataset, we ran K-means, bisect K-means, NMF and NCW-NMF 50 times with different initial values. We then obtained the result of EN-NCW-NMF by aggregating 50 clustering result of NCW-NMF. As shown in Table 2, we present the clustering result of these 7 different algorithms. For each given cluster number k, the performance of each algorithm is averaged across all 10 datasets, and the highest NMI is highlighted with bold face.

Without considering ensemble methods, we first compare Hierarchical Clustering, K-means, bisect K-means, NMF and NCW-NMF. The experimental results show that NCW-NMF outperform all other methods significantly in almost all cases with different k. The average NMI achieved by NCW-NMF is 0.7760, which is followed by bisect K-means(0.7272), K-means(0.7147), NMF(0.6867) and Hierarchical clustering(0.5435). For a specific k, NCW-NMF also surpasses all other methods except $k = 8$. Bisect K-means, K-means and NMF achieve close performance, while bisect K-means outperforms the other two slightly. In our experiment, we found that Hierarchical clustering performs worst, which is consist with Yoo et al's previous comparison result[12]. Furthermore, by aggregating clustering result of different runs, EN-NCW-NMF obtained an NMI value of 0.7588, which is the second highest value achieved by all algorithms. Although NCW-NMF achieves slightly higher NMI than EN-NCW-NMF, the stochastic nature of NCW-NMF makes its performance very sensitive to initialization value. For example, Min-NCW-NMF, which represent the worst case of NCW-MMF in each round, only achieves an average NMI of 0.6706. The under-performed Min-NCW-NMF justifies the reasonability of EN-NCW-NMF, which can achieve robust and effective clustering results.

[1] http://glaros.dtc.umn.edu/gkhome/views/cluto

Table 2: Performance comparisons (NMI) using Genomics2004 collection

k	Hierarchical Clustering	K-means	Bisecting K-Means	NMF	NCW-NMF	Min-NCW-NMF	Ensemble NCW-NMF
3	0.4670	0.5520	0.5735	0.5402	**0.6372**	0.5294	0.6294
4	0.4694	0.6870	0.6892	0.7002	**0.7969**	0.6399	0.7736
5	0.6013	0.7391	0.7384	0.7347	**0.8331**	0.7071	0.8123
6	0.5233	0.6991	0.7081	0.6744	**0.7625**	0.6623	0.7383
7	0.5427	0.7527	0.7705	0.7102	**0.7866**	0.6733	0.7818
8	0.6369	0.7837	**0.8033**	0.7253	0.8016	0.7191	0.7998
9	0.5514	0.7577	0.7720	0.7022	**0.7956**	0.7329	0.7632
10	0.5563	0.7463	0.7629	0.7064	**0.7942**	0.7302	0.772
Mean	0.5435	0.7147	0.7272	0.6808	**0.7760**	0.6743	0.7588

Furthermore, we compared the result of EN-NCW-NMF with bisect K-means over each dataset. Assume the distribution of NMI using bisect K-means is a normal distribution, we use Z-value to measure the significance of improvement. The result shows that, EN-NCW-NMF achieved a higher NMI than average of Bisect K-Means on 54 out of all 80 datasets, and on 38 datasets of them, EN-NCW-NMF achieves an Z-value of 1.96 or higher (97.5% significant). We can clearly see that EN-NCW-NMF can obtain not only robust but also accurate clustering result by aggregating individual NMF results with different initialization parameters. In this work, MCLA is used in the integration phase of ensemble clustering. Other integration algorithms can be also incorporated to improve the clustering performance.

4 Conclusion and Future Work

In this work we propose to use ensemble NMF to clustering biomedical documents. Although being a popular dimension reduction technique, NMF suffers from its stochastic nature, and thus the performance is sensitive to the initialization parameters. Through aggregating various clustering result of different runs, ensemble NMF can produce very good clustering result robustly. Future work includes developing suitable aggregation algorithms and examining our algorithms on large scale datasets.

References

[1] J. Ghosh. Scalable clustering methods for data mining. In N. Ye, editor, *Handbook of data mining*. Lawrence Erlbaum, 2003.

[2] WR. Hersh, RT. Bhupatiraju, L. Ross, P. Johnson, AM. Cohen, and DF. Kraemer. TREC 2004 genomics track overview. In E. M. Voorhees and Lori P. Buckland, editors, *the proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.

[3] LJ. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7(2):119–129, 2006.

[4] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, pages 788–791, 1999.

[5] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems,*, pages 556–562, 2001.

[6] M. Lee, W. Wang, and H. Yu. Exploring supervised and unsupervsied methods to detect topics in biomedical text. *BMC Bioinformatics*, page 140, 2006.

[7] C.-J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, pages 2756–2779, 2007.

[8] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR)*, 3:583–617, December 2002.

[9] SA. Vavasis. On the complexity of nonnegative matrix factorization. `http://arxiv.org/abs/0708.4149`, 2007.

[10] D. Wheeler et al. Database resources of the national center for biotechnology information. *Nucl. Acids Res.*, 33:D39–D45, 2005.

[11] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix fatorization. In *Proceedings of SIGIR2003 July 28- Aug 1, Toronto, Canada*, pages 267–273, 2003.

[12] I. Yoo and X. Hu. A comprehensive comparison study of document clustering for a biomedical digital library medline. In Gary Marchionini, Michael L. Nelson, and Catherine C. Marshall, editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006, Chapel Hill, NC, USA, June 11-15, 2006, Proceedings*, pages 220–229, 2006.

[13] S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4:1001–1037, 2003.

[14] S. Zhong and J. Ghosh. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3):374–384, 2005.