LEHIGH UNIVERSITY

CSE 441 PROJECT REPORT

Document Clustering via Non-negative Matrix Factorization

Author:
Ziyi GUO

Lecturer:
Prof. Henry BAIRD

April 23, 2013

Document Clustering via Non-negative Matrix Factorization

Ziyi Guo
Department of Computer Science and Engineering
Lehigh University
zig312@lehigh.edu

April 23, 2013

Abstract

In the field of pattern recognition and machine learning, how to identify data patterns is always a challenging problem and has important implication in further research and real-world applications, such as search engine and recommendation systems. Non-negative matrix factorization (NMF), a part-based, linear combination, unsupervised learning algorithm, was shown to be useful in analyzing and interpreting large-scale data. In this project, we apply NMF as a method of dimension reduction on the task of document clustering. In the latent space derived by NMF, each axis captures one topic of a particular document cluster, and each document is represented as a linear combination of all the topic clusters, and K-means algorithm is used to determine the topic label of each document. The experiments on three benchmark document dataset show that the proposed NMF-based clustering method is capable of identifying latent structures in the documents, and outperforms the traditional PCA-based method.

I Introduction

In the field of data mining and pattern recognition, clustering analysis refers to the division of data into several groups. Each group, also called cluster, contains data points that are similar with themselves and dissimilar with points in other groups. Similarly, the task of document clustering is grouping a set of documents (e.g. web pages, online news) into clusters or topics. In recent years, document clustering has received more and more attentions since it is a fundamental task in the field of data mining because it enables useful tools for efficient organization, navigation and retrieval of huge amount of text documents, being great important to further research and real-world applications, such as search engine and recommendation systems. For example, in news recommendation systems, the key challenge is to help users find news articles that are interesting to read. Therefore, clustering news documents according to their topics is an indispensable part since most readers are just interested in news of specific topics. One online news recommendation example is shown in Figure 1.

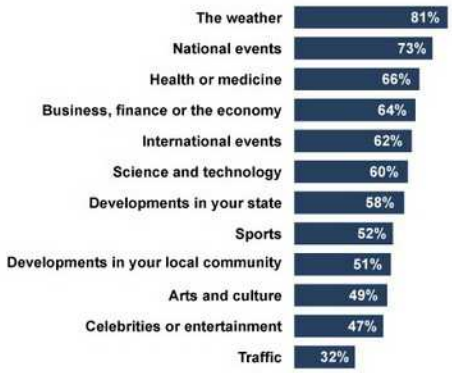

Figure 1: News Recommendation Topics. The number represents the percentage of interested users for each topic.

Besides, it is very important to distinguish document clustering (unsupervised learning) from document classification (supervised learning). In supervised classification task, we are given a set of labeled text documents, and our task is to design advanced algorithms to identify patterns in these document and use these patterns to predict the topics of newly encountered, unlabeled documents. However, in the case of unsupervised clustering problem, no prior label information is provided, and thus the problem is to group a given dataset into meaningful clusters.

II Related Works

Document clustering methods, or data clustering in general, can be mainly categorized into two types, agglomerative methods and partitioning methods [1]. Agglomerative methods group all the data points, each of which represents one document into a hierarchical tree structure, or a dendrogram via the bottom-up approach. Beginning from each data point in a distinct cluster, we iteratively merge those two points with the least distance into one cluster in each step. Upon completion, all the points are clustered into one group and the cluster label of each data point can be derived from the tree structure. One example of agglomerative clustering is shown in Figure 2.

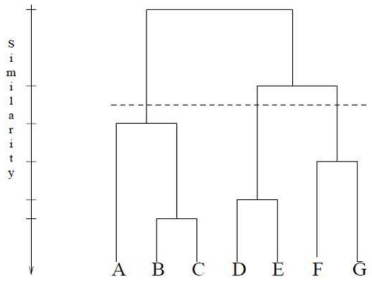

Figure 2: Agglomerative Clustering Example [1].

The computational complexity of such clustering methods is often $O(n^2 \log n)$ and n represents the number of documents in the dataset. Due to the quadratic order of complexity, the agglomerative methods are usually computationally prohibitive when dealing with big data in the real world.

In the case of partitioning methods, we decompose all the data points into a given number of disjoint clusters. Then, we compute the distance between each point and all the cluster center, label each point into the nearest cluster and update the center position for each cluster. One example of partitioning clustering is shown in Figure 3.

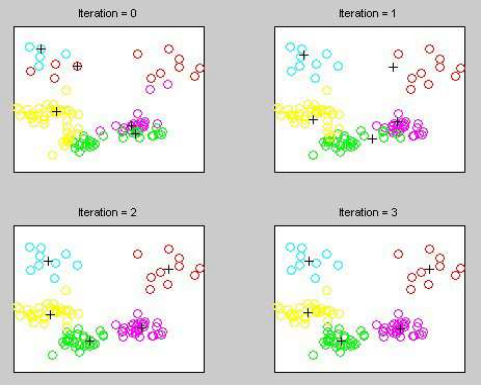

Figure 3: Partitioning Clustering Example [1]. The color indicates the label of each cluster and the clustering results are shown in each iteration.

However, in the real world applications, the dimension of data points in the feature space is very high, usually in the thousands or millions, leading to the curse of dimensionality, which is defined as various phenomena when analyzing and organizing data points in high-dimensional space that cannot be found in lower dimensions. Therefore, to solve the problem of the curse of dimensionality, the technique of dimension reduction is often applied in document partitioning clustering.

Principal component analysis (PCA), which is a process of converting a set of observations of correlated variables into a set of values of linearly uncorrelated variables, is one of the most well-known methods for dimension reduction. However, PCA will create negative values in the decomposed vectors, breaking the nonnegative constraints of document-term matrix in document clustering. To deal with such constraints, the nonnegative matrix factorization (NMF) [2] method has been proposed. Xu&Liu (2003) [3] and Shahnaz&Berry (2004) [4] have showed that NMF is capable of identifying latent structures and outperformed other rank reduction methods in document clustering. In this project, we experimentally verify the use of NMF in document partitioning clustering. First, we briefly introduce NMF along with its multiplicative update algorithm; then we introduce the NMF-based document clustering; finally, we test the proposed method on three benchmark document dataset and compare it with the traditional PCA-based method.

III NMF Algorithm

NMF is a linear, non-negative approximate data representation [2, 5]. Given a $N \times M$ non-negative matrix V, where each column represents one data point in the N dimensional space, we find two non-negative decomposed matrices W and H such that:

$$V \approx WH \tag{1}$$

Where W has size $N \times k$ and H has size $k \times M$ and k is much smaller than N and M or:

$$v \approx Wh \tag{2}$$

Where v and h are the corresponding columns of V and H. That is to say, each original data point can be approximated by a linear combination of the columns in W weighted by the corresponding column in H. Therefore, W can be viewed as the basis matrix and H can be viewed as the coefficient matrix, as shown in Figure 4.

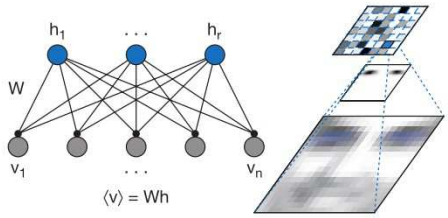

Figure 4: The Underlying Model of NMF [2]. Each visible output v_i can be generated by the linear combination of all the latent variable h_i and the right sub figure is a facial representation example.

To find an approximate factorization for $V \approx WH$, we should first define the cost function between the original matrix V and the factorized matrices multiplication textitWH. Typically, the square of the Euclidean distance is applied:

$$||A-B||^2 = \sum_{ij} (A_{ij} - B_{ij})^2 \tag{3}$$

Then, the Euclidean distance is nonincreasing under the update rules and the cost function converges to a local minima [5]:

$$\begin{aligned} H_{au} &\leftarrow H_{au} \frac{(W^T V)_{au}}{(W^T WH)_{au}} \\ W_{ia} &\leftarrow W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \end{aligned} \tag{4}$$

Given the input V and two random initial matrices W and H, iterate the update process via multiplicative rules until convergence or reaching maximum number of iterations and we obtain the final decomposed factors

.

NMF has a variety of advantages. 1) The non-negative constraint of the input matrix is required in many real-world problems, such as the word frequency in a given document, the pixel values in one image and the gene expression level of one sample. These non-negative constraints cannot be directly explained by other matrix factorization methods, such as PCA and SVD. 2) In NMF, only additive combinations are allowed, and this part-based representation corresponds to the fact that the whole is made of all parts. 3) NMF is fast for convergence. Here, we show the relationship between the NMF object function and the number of iterations on the input matrix with size of 1033×4328. From the figure above, we find that the algorithm converged using about 25 itera-

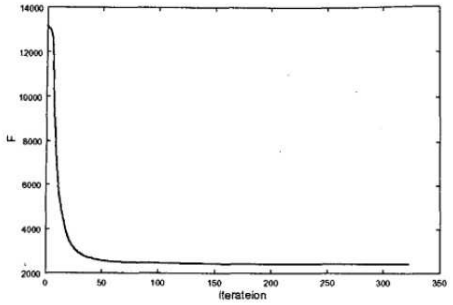

Figure 5: Cost of objective function as a function of number of iterations [6].

tions, which can be implemented in most PCs. With these advantages, NMF has been widely used in data mining [3, 4], computer vision [7], natural language processing [8] and bioinformatics [9]. In next section, we talk about the use of NMF in document clustering.

IV NMF-based Document Clustering

Using the vector space model [10], the document dataset can be expressed as a $m \times n$ matrix V, where m is the number of terms in the dictionary and n is the number of documents in the dataset. Therefore, each column of V encodes one document and each entry v_{ij} is the significance of word i with respect to the document j. Here, each document is represented as a weighted term-frequency vector after the removal of stop words and word stemming. Then, we recalculate the weight of each feature via the term frequency-inverse document frequency (tf-idf):

$$v_{ij} = tf_{ij} \times \log \frac{n}{idf_j} \tag{5}$$

Where v_{ij}, tf_{ij}, n, idf_j denotes the term frequency-inverse document frequency of word i in document j, the term frequency of word i in document j, the number of total documents, and the number of document containing word i respectively. Next, using the multiplicative update rules, the term-document matrix V is decomposed into the $m \times k$ basis matrix W and the $k \times n$ coefficient matrix H. Each entry of H, h_{ij} ,indicates the significance of the latent topic i in the document j, so H can be viewed as the latent low-rank representation of the original matrix V. In the work of Xu&Liu(2003), the topic of each document is assigned to cluster x if $x = \underset{j}{argmax}\, h_{ij}$. However,

in this project, we take H as the latent reduced space and apply K-means on this space for the clustering.

In summary, the NMF-based document clustering is composed of the following steps:
1. Do the pre-processing on the original document set and build the term-document matrix V.
2. Take the term frequency-inverse document frequency strategy on V.
3. Perform NMF on V and get the decomposed factors W and H.
4. Use K-means clustering on the latent reduced space H and return cluster labels for all the documents.

V Performance Evaluation

V.I Data Corpora

We conduct the performance evaluations on three benchmark text document corpora: Top 30 topics in TDT2, 20 Newsgroups and Reuters21578. 1) The TDT2 corpus (Nist Topic Detection and Tracking Corpus) is composed of news documents during the first half of 1998 from 6 media sources (APW, NYT, VOA, PRI, CNN and ABC). It consists of 11201 documents in 96 distinct semantic categories. In this project, the documents in more than one category are removed and the largest 30 topics are used, leaving 9394 documents and 36771 words. 2) The Reuters21578 dataset consists of collection on the Reuters newswire in 1987, and the documents were assembled and indexed with topics manually in 1987. This corpus contains 21578 documents in 135 categories. Also, the documents in more than one category are removed, leaving 8293 documents in 65 categories and 18933 words. 3) The 20 Newsgroups dataset is a collection of about 20000 news documents and the topics range from computer, science to political talks. The dataset consists of 18846 documents in 20 topics.

V.II Evaluation Metrics

Two metrics, the accuracy (AC) and the normalized mutual information ($\bar{MI}$), are used to evaluate the clustering performance. Given a document d_i, the clustering label l_i and the ground truth label α_i, the AC is defined as:

$$AC = \frac{\sum_{i=1}^{n} \delta(\alpha_i, map(l_i))}{n} \tag{6}$$

Where n is the number of documents in the dataset, $\delta(x,y)$ is the delta function that equals one if $x = y$) and zero otherwise, and map is the function that best maps each clustering label to the corresponding ground truth label. Given two document clusters C and C', the mutual information MI is defined as:

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_i)} \tag{7}$$

Where $p(c_i)$ and $p(c'_i)$ indicate the probabilities that an arbitrarily selected document belongs to the clusters C and C' respectively, and $p(c_i, c'_j)$ indicates the probability that an arbitrarily selected

document belongs to both C and C'. The value of $MI(C,C')$ is between zero and $max(H(C),H(C'))$ where $H(C)$ and $H(C')$ are the entropies of C and C' respectively. To simplify the MI values, the normalized mutual information $\bar{MI}$ whose value is between zero and one is defined as:

$$\bar{MI}(C,C') = \frac{MI(C,C')}{max(H(C),H(C'))} \tag{8}$$

V.III Evaluation Results

In this section, we compare the NMF-based document clustering method with K-means without dimension reduction and PCA-based K-means method on three document corpora. The results are shown in Table 1.

	Top30 TDT2			Reuters21578			20 Newsgroup		
	K-means	PCA+K-means	NMF+K-means	K-means	PCA+K-means	NMF+K-means	K-means	PCA+K-means	NMF+K-means
AC	0.60709	0.58452	0.7193	0.24683	0.28422	0.23369	0.4836	0.43171	0.49224
MI	0.72195	0.71968	0.7741	0.37797	0.38059	0.35699	0.48362	0.45819	0.49257

Table 1: Clustering Performance Comparisons on Document Dataset

Also, we compare these three algorithms on two benchmark image libraries. The first one is COIL20 image library, which contains 32×32 gray scale images of 20 objects from different angles. The second one is the CMU PIE face library, which contains 32×32 gray scale face images of 68 peoples. Each person has 21 facial images under different light conditions. The data clustering on images is almost identical to document clustering except that the feature of each image represent the pixel values and each column of the input matrix V is normalized rather than taking tfidf weighting. The clustering results are shown in Table 2.

	COIL20 Face			PIE Pose		
	K-means	PCA+K-means	NMF+K-means	K-means	PCA+K-means	NMF+K-means
AC	0.50833	0.52917	0.60972	0.22934	0.2423	0.41001
MI	0.67132	0.67398	0.70928	0.47765	0.48393	0.70124

Table 2: Clustering Performance Comparisons on Image Dataset

The experimental results show that: 1) NMF outperforms K-means and PCA+K-means on four out of five datasets. This indicates that NMF learns a better compact representation in the sense of semantic latent structure. 2) NMF fails to get a better result than K-means and PCA+K-means on the Reuters21578 dataset. The possible reason maybe that there are too many latent topics, and this indicates that the complexity of data semantic structure may influence the performance of NMF.

VI Discussion

In summary, NMF is an emerging new algorithm for large-scale data analysis and interpretation. In this project, we apply NMF as a method of dimension reduction on document clustering. The experiments show that NMF is capable of identifying latent structures in the big data and outperforms PCA-based method. Actually, some extensions of NMF, such as Sparse Non-negative Matrix Factorization (SNMF) [11] and Graph-regularized Non-negative Matrix Factorization (GNMF) [12],

have been proposed. Their performances in data mining and other machine learning related areas still need more efforts.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[3] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

[4] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, no. 2, pp. 373–386, 2006.

[5] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, pp. 556–562, 2001.

[6] S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kita, "Dimensionality reduction using non-negative matrix factorization for information retrieval," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*. IEEE, 2001, pp. 960–965.

[7] W. Liu and N. Zheng, "Non-negative matrix factorization based methods for object recognition," *Pattern Recognition Letters*, no. 8, pp. 893–897, 2004.

[8] T. Li, Y. Zhang, and V. Sindhwani, "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*. Association for Computational Linguistics, 2009, pp. 244–252.

[9] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS computational biology*, vol. 4, no. 7, 2008.

[10] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[11] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[12] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 63–72.