Preprocessing using Non-negative Matrix Factorization in Conjunction with K-means

Ralph Abbey,  Carl Meyer

NCSU

SIAM-SEAS: 21st, March 2010
1. Introduction

2. Document Clustering
   - Document Data
   - Term by Document Matrix

3. Clustering Methods
   - K-means
   - Non-negative Matrix Factorization (NMF)

4. NMF as a preprocessor with K-Means

5. Results

6. Concluding Remarks
Introduction

Why clustering? Who does this?

Is there one clustering method that is better than others?

How does this affect me?
Introduction

Why clustering? Who does this?
Introduction

- Why clustering? Who does this?
- Is there one clustering method that is better than others?
Introduction

- Why clustering? Who does this?
- Is there one clustering method that is better than others?
- How does this affect me?
Kendall and Babcock defined an object $A$ is preferred to an object $B$ in a given set $D$ of $n$ objects and in a complete $n$-tournaments if the number of preferences $\xi$ is a circular trials, i.e., $A \rightarrow B$, $B \rightarrow C$, and $C \rightarrow A$. Once a complete set of preferences was defined to depend on the number of trials among the preferences [36]. $\xi = 1$ if no ties among the preferences. $\xi$ decreases to 0 as the number complete set of preferences increases. The number of circular trials, $c$, can also be interpreted as the number of preference reversals necessary to break all ties in the score vector $\mathbf{a}$ (i.e., the number of times $i$ is preferred to other objects). Once all the ties are removed the complete set of preferences represents a ranking, also called a transitive $n$-tournament [25], or a linear ordering [26] that is not necessarily unique. David calls the resulting ranking a nearest adjoing order. Slater proposed a different measure of inconsistency he called $d$ that is the minimum number of preference reversals needed to reach a nearest adjoing order. note that $1 \leq c \leq |2d|$. Another type of inconsistency has been studied by Gehr and Shapiron [47]. If a prior ordering of the objects has $A \rightarrow B \rightarrow C$, Gehr and Shapiron call the situation in which the

**Figure:** A pdf document

**Figure:** An email
Term by Document Matrix (TBD)

The element $A_{i,j}$ counts the number of times word $i$ appears in document $j$.

Consider the example with 3 documents:

- **Document 1** has the words "apple" twice, "bear" once, "cannon" four times.
- **Document 2** has the words "bear" three times, "cannon" once, and "disco" once.
- **Document 3** has the words "apple" 5 times, and "disco" twice.

\[
A = \begin{bmatrix}
2 & 0 & 5 \\
1 & 3 & 0 \\
4 & 1 & 0 \\
0 & 1 & 2 \\
\end{bmatrix}
\]
Term by Document Matrix (TBD)

- The element $A_{i,j}$ counts the number of times word $i$ appears in document $j$.
The element $A_{i,j}$ counts the number of times word $i$ appears in document $j$

Consider the example with 3 documents:

- document 1 has the words “apple” twice, “bear” once, “cannon” four times
- document 2 has the words “bear” three times, “cannon” once, and “disco” once
- document 3 has the words “apple” 5 times, and “disco” twice.

$$TBD = \begin{pmatrix} 2 & 0 & 5 \\ 1 & 3 & 0 \\ 4 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$$
What is K-means?

The goal of K-means clustering is to minimize the sum of squared distances between each data point and its assigned centroid. This is mathematically expressed as:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} (d_i - c_j)^2$$

where $d_i$ represents a data point, $c_j$ is a centroid, and $k$ is the number of clusters.

The iterative process of K-means involves:
1. Assigning each data point to the nearest centroid in the Euclidean sense.
2. Recalculating the centroids by finding the average of all data points assigned to each cluster, i.e.,
   $$c_j = \frac{1}{L_j} \sum_{i=1}^{L_j} d_i$$
   where $L_j$ is the number of data points assigned to cluster $j$.
3. Repeating the process until convergence to a local minimum is achieved.

Preprocessing using Non-negative Matrix Factorization in conjunction with K-means is discussed in the context of the presentation.
What is K-means?

Goal is to minimize:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} (d_i - c_j)^2$$
What is K-means?

- Goal is to minimize:
  \[
  \sum_{i=1}^{n} \sum_{j=1}^{k} (d_i - c_j)^2
  \]

- Iterative process in which iterations continue until convergence to a local minimum
What is K-means?

- Goal is to minimize:

\[
\sum_{i=1}^{n} \sum_{j=1}^{k} (d_i - c_j)^2
\]

- Iterative process in which iterations continue until convergence to a local minimum

- At each step: assign documents to the centroid to which they are closest to in the Euclidean sense
What is K-means?

- Goal is to minimize:
  \[ \sum_{i=1}^{n} \sum_{j=1}^{k} (d_i - c_j)^2 \]

- Iterative process in which iterations continue until convergence to a local minimum

- At each step: assign documents to the centroid to which they are closest to in the Euclidean sense

- Then recalculate centroids by finding the average of all documents assigned to the centroid, that is: 
  \[ c_j = \frac{\sum_{i=1}^{L} d_i}{L} \]
  where L is the number of documents assigned to cluster j, and the division is a scalar division of the elements of d.
What is the Non-negative Matrix Factorization?
What is the Non-negative Matrix Factorization?

\[ A_{m \times n} \approx W_{m \times r} H_{r \times n}, \quad A, W, H \geq 0, \quad r \in N \] is user defined
**What is the Non-negative Matrix Factorization?**

- $A_{m \times n} \approx W_{m \times r}H_{r \times n}$, $A, W, H \geq 0$, $r \in N$ is user defined
- The goal is to minimize $\|A - WH\|$
What is the Non-negative Matrix Factorization?

- $A_{m \times n} \approx W_{m \times r} H_{r \times n}$, $A, W, H \geq 0$, $r \in \mathbb{N}$ is user defined
- The goal is to minimize $\|A - WH\|$
- A class of algorithms - not just one
An Algorithm for the NMF

An Algorithm for the NMF

Lee and Seung 1999

Iteratively update until the error $\|A - WH\|_F$ is below some threshold.

$H_{i,j} = H_{i,j}(W^T A)_{i,j}(W^T W)_{i,j} + \epsilon$

$W_{i,j} = W_{i,j}(A H^T)_{i,j}(W^T H^T)_{i,j} + \epsilon$

Guaranteed convergence to a local min
An Algorithm for the NMF

Lee and Seung 1999
An Algorithm for the NMF

- Lee and Seung 1999
- Iteratively update until the error \( \|A - WH\|_F^2 \) is below some threshold.
Clustering Methods

Non-negative Matrix Factorization (NMF)

An Algorithm for the NMF

- Lee and Seung 1999
- Iteratively update until the error $\|A - WH\|_F^2$ is below some threshold.

\[
H_{i,j} = H_{i,j} \frac{(W^T A)_{i,j}}{(W^T WH)_{i,j} + \epsilon}
\]

\[
W_{i,j} = W_{i,j} \frac{(AH^T)_{i,j}}{(WHH^T)_{i,j} + \epsilon}
\]
An Algorithm for the NMF

- Lee and Seung 1999
- Iteratively update until the error $\|A - WH\|_F^2$ is below some threshold.

\[
H_{i,j} = H_{i,j} \frac{(W^T A)_{i,j}}{(W^T WH)_{i,j} + \epsilon}
\]

\[
W_{i,j} = W_{i,j} \frac{(AH^T)_{i,j}}{(WHH^T)_{i,j} + \epsilon}
\]

- Guaranteed convergence to a local min
NMF used in Clustering

Remember, we are looking at \( A_{m \times n} \approx W_{m \times r}H_{r \times n} \)
NMF used in Clustering

- Remember, we are looking at $A_{m \times n} \approx W_{m \times r} H_{r \times n}$
- $\hat{a}_j = \sum_{i=1}^{r} h_{i,j} w_i$ The coefficients in $H$ are (approximately) the coordinates of the data points with respect to the basis for the feature space.
NMF used in Clustering

- Remember, we are looking at $A_{m \times n} \approx W_{m \times r} H_{r \times n}$
- $\hat{a}_j = \sum_{i=1}^{r} h_{i,j} w_i$ The coefficients in $H$ are (approximately) the coordinates of the data points with respect to the basis for the feature space.
- The standard method of clustering using the NMF is done by setting $r = k$, where $k$ is the number of clusters desired.
Clustering Methods

Non-negative Matrix Factorization (NMF)

NMF used in Clustering

- Remember, we are looking at $A_{m \times n} \approx W_{m \times r} H_{r \times n}$
- $\hat{a}_j = \sum_{i=1}^{r} h_{i,j} w_i$ The coefficients in $H$ are (approximately) the coordinates of the data points with respect to the basis for the feature space.
- The standard method of clustering using the NMF is done by setting $r = k$, where $k$ is the number of clusters desired.
- The clustering is then computed by associating document $i$ with cluster $j$ if the $j$th element in column $i$ of $H$ is the maximum entry in that column.
The coefficients in $H$ are (approximately) the coordinates of the data points with respect to the basis for the feature space. Thus we can treat $H$ as a "new" TBD, in which the "terms" are really the columns of $W$. We call $W$ the "feature basis", as it has picked out features to be the new terms in $H$. Now we can cluster $H$. There is no restriction on the $r$ we choose for the NMF, but observation has shown that $r \approx 3k$ works well.
The coefficients in $H$ are (approximately) the coordinates of the data points with respect to the basis for the feature space.
The coefficients in $H$ are (approximately) the coordinates of the data points with respect to the basis for the feature space.

Thus we can treat $H$ as a “new” TBD, in which the “terms” are really the columns of $W$. We call $W$ the “feature basis”, as it has picked out features to be the new terms in $H$. 
The coefficients in $H$ are (approximately) the coordinates of the data points with respect to the basis for the feature space.

Thus we can treat $H$ as a “new” TBD, in which the “terms” are really the columns of $W$. We call $W$ the “feature basis”, as it has picked out features to be the new terms in $H$.

Now we can cluster $H$. There is no restriction on the $r$ we choose for the NMF, but observation has shown that $r \approx 3k$ works well.
Benchmark Document Sets

Used Medline, Cranfield, Cisi datasets, with 1033, 1460, and 1398 documents respectively.

Combined the three document sets into one overall set, and then clustered with $k=3$ to try to recover the original separated sets.

The metric for determining cluster quality was an accuracy metric $\sum_{k=1}^{3} \frac{\# \text{correctly clustered}}{\# \text{total}}$ - can think of as a percent correct.

Ralph Abbey, Carl Meyer (NCSU)
Benchmark Document Sets

- Used Medline, Cranfield, Cisi datasets, with 1033, 1460, and 1398 documents respectively
Benchmark Document Sets

- Used Medline, Cranfield, Cisi datasets, with 1033, 1460, and 1398 documents respectively
- Combined the three document sets into one overall set, and then clustered with $k = 3$ to try to recover the original separated sets
Benchmark Document Sets

- Used Medline, Cranfield, Cisi datasets, with 1033, 1460, and 1398 documents respectively.
- Combined the three document sets into one overall set, and then clustered with $k = 3$ to try to recover the original separated sets.
- The metric for determining cluster quality was an accuracy metric:
  $$\sum_{i=1}^{k} \frac{\# \text{correctly clustered}}{\text{total}\#}$$
  - can think of as a percent correct.
### Results

Each were run 200 times

**Table:** Results of k-means, and nmf preprocessing to k-means

<table>
<thead>
<tr>
<th></th>
<th>k-means</th>
<th>nmf</th>
<th>$r = 6$</th>
<th>$r = 9$</th>
<th>$r = 12$</th>
</tr>
</thead>
<tbody>
<tr>
<td>min. acc.</td>
<td>0.586</td>
<td>0.465</td>
<td>0.493</td>
<td>0.498</td>
<td>0.523</td>
</tr>
<tr>
<td>max acc.</td>
<td>0.886</td>
<td>0.957</td>
<td>0.962</td>
<td>0.965</td>
<td>0.965</td>
</tr>
<tr>
<td>avg. acc.</td>
<td>0.727</td>
<td>0.623</td>
<td>0.766</td>
<td>0.771</td>
<td>0.755</td>
</tr>
<tr>
<td>var. acc.</td>
<td>0.0077</td>
<td>0.0055</td>
<td>0.0269</td>
<td>0.0285</td>
<td>0.0251</td>
</tr>
</tbody>
</table>
Figure: Methods of clustering with means and 95% confidence intervals
Concluding Remarks

K-means and NMF work well on their own, but work better together.

NMF has already been used for preprocessing in information retrieval.

Further areas of research:
- Apply this method to other areas aside from document clustering
- Try other clustering algorithms along with NMF preprocessing

Ralph Abbey, Carl Meyer (NCSU)
K-means and NMF work well on their own, but work better together.
Concluding Remarks

- K-means and NMF work well on their own, but work better together
- NMF has already been used for preprocessing in information retrieval
Concluding Remarks

- K-means and NMF work well on their own, but work better together
- NMF has already been used for preprocessing in information retrieval
- Further areas of research:
  - Apply this method to other areas aside from document clustering
  - Try other clustering algorithms along with NMF preprocessing