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

Introduction

- Why clustering? Who does this?

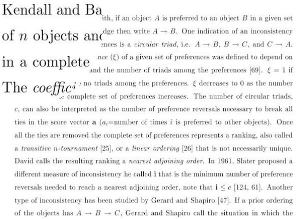
Introduction

- 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?

What Does Document Data Look Like?



Kendall and Ba ith, if an object A is preferred to an object B in a given set

of n objects anc dge then write $A \to B$. One indication of an inconsistency

ences is a *circular triad*, i.e. $A \to B$, $B \to C$, and $C \to A$.

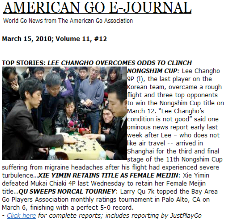
in a complete nce (ξ) of a given set of preferences was defined to depend on

and the number of triads among the preferences [69]. $\xi = 1$ if

The *coeffici* no triads among the preferences. ξ decreases to 0 as the number

complete set of preferences increases. The number of circular triads, c, can also be interpreted as the number of preference reversals necessary to break all ties in the score vector $\mathbf{a}$ (a_i=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 adjoining order*. In 1961, Slater proposed a different measure of inconsistency he called $\mathbf{i}$ that is the minimum number of preference reversals needed to reach a nearest adjoining order, note that $\mathbf{i} \leq c$ [124, 61]. Another type of inconsistency has been studied by Gerard and Shapiro [47]. If a prior ordering of the objects has $A \to B \to C$, Gerard and Shapiro call the situation in which the

Figure: A pdf document



AMERICAN GO E-JOURNAL

World Go News from The American Go Association

March 15, 2010; Volume 11, #12

TOP STORIES: *LEE CHANGHO OVERCOMES ODDS TO CLINCH NONGSHIM CUP*: Lee Changho 9P (l), the last player on the Korean team, overcame a rough flight and three top opponents to win the Nongshim Cup title on March 12. "Lee Changho's condition is not good" said one ominous news report early last week after Lee – who does not like air travel -- arrived in Shanghai for the third and final stage of the 11th Nongshim Cup suffering from migraine headaches after his flight had experienced severe turbulence...***XIE YIMIN RETAINS TITLE AS FEMALE MEIJIN***: Xie Yimin defeated Mukai Chiaki 4P last Wednesday to retain her Female Meijin title...***QU SWEEPS NORCAL TOURNEY***: Larry Qu 7k topped the Bay Area Go Players Association monthly ratings tournament in Palo Alto, CA on March 6, finishing with a perfect 5-0 record.

- *Click here for complete reports; includes reporting by JustPlayGo*

Figure: An email

Term by Document Matrix (TBD)

Term by Document Matrix (TBD)

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

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.

$$TBD = \begin{pmatrix} 2 & 0 & 5 \\ 1 & 3 & 0 \\ 4 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$$

What is K-means?

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: $c_j = \sum_{i=1}^{L} \frac{d_i}{L}$, 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?

- $A_{m \times n} \approx W_{m \times r} H_{r \times n}$, $A, W, H \geq 0$, $r \in N$ 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?

- $A_{m\times n} \approx W_{m\times r}H_{r\times n}$, $A, W, H \geq 0$, $r \in N$ 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

An Algorithm for the NMF

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

An Algorithm for the NMF

- Lee and Seung 1999
- Iteratively update until the error $\|A - WH\|_F^2$ is below some threshold.
- $$H_{i,j} = H_{i,j}\frac{(W^TA)_{i,j}}{(W^TWH)_{i,j} + \epsilon}$$ $$W_{i,j} = W_{i,j}\frac{(AH^T)_{i,j}}{(WHH^T)_{i,j} + \epsilon}$$

An Algorithm for the NMF

- Lee and Seung 1999
- Iteratively update until the error $\|A - WH\|_F^2$ is below some threshold.
- $$H_{i,j} = H_{i,j}\frac{(W^TA)_{i,j}}{(W^TWH)_{i,j} + \epsilon}$$
  $$W_{i,j} = W_{i,j}\frac{(AH^T)_{i,j}}{(WHH^T)_{i,j} + \epsilon}$$
- 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 $A_{m\times n} \approx W_{m\times r} H_{r\times n}$
- $\hat{\mathbf{a}}_j = \sum_{i=1}^{r} \mathbf{h}_{i,j}\mathbf{w}_i$ 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 $A_{m\times n} \approx W_{m\times r} H_{r\times n}$
- $\hat{\mathbf{a}}_j = \sum_{i=1}^{r} \mathbf{h}_{i,j}\mathbf{w}_i$ 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.

NMF used in Clustering

- Remember, we are looking at $A_{m\times n} \approx W_{m\times r} H_{r\times n}$
- $\hat{\mathbf{a}}_j = \sum_{i=1}^{r} \mathbf{h}_{i,j}\mathbf{w}_i$ 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 jth element in column i of H is the maximum entry in that column.

NMF preprocessing

NMF preprocessing

- The coefficients in H are (approximately) the coordinates of the data points with respect to the basis for the feature space.

NMF preprocessing

- The coefficients in H are (approximately) the coordinates of the data points with respect to the basis for the feature space.
- Thus we can treate 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.

NMF preprocessing

- The coefficients in H are (approximately) the coordinates of the data points with respect to the basis for the feature space.
- Thus we can treate 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

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 $\sum_{i=1}^{k} \frac{\#correctly\_clustered}{total\#}$ - can think of as a percent correct

Results

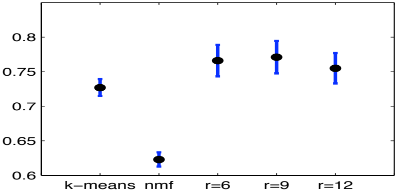
Each were run 200 times

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

	k-means	nmf	$r = 6$	$r = 9$	$r = 12$
min. acc.	0.586	0.465	0.493	0.498	0.523
max acc.	0.886	0.957	0.962	0.965	0.965
avg. acc.	0.727	0.623	0.766	0.771	0.755
var. acc.	0.0077	0.0055	0.0269	0.0285	0.0251

Results continued

Figure: Methods of clustering with means and 95% confidence intervals



Concluding Remarks

Concluding Remarks

- 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