A Comparison of Methods for Creating Similarity Matrices

Carl D. Meyer
Charles D. Wessel

North Carolina State University

34th SIAM Southeastern-Atlantic Section Conference
March 21, 2010

1 Introduction

2 From vectors to similarity

3 The Running Example

4 Clustering Similarity Matrices

5 Clustering the Running Example

6 Results from another data set

The Big Idea

When clustering a data set with a particular algorithm it is important that

- we have an idea of what we mean by *similarity*.
- the algorithm used meshes with this idea.
- the similarity measure is appropriate for the data being clustered.

What is a cluster?

- A group of *objects* from a data set
- Examples of *objects*: text documents, visual images, medical samples, movies, movie critics, voting precincts
- Grouped objects are *similar* in some way
- Ungrouped objects are *dissimilar* in some way
- The notion of *similarity* is very important

The Fundamental Theorem of Applied Mathematics

The Fundamental Theorem of Applied Mathematics

Nothing works.

The Fundamental Theorem of Cluster Analysis

Theorem. *There is no best clustering method, that is, one which is superior to all other methods for solving all problems in a particular class of problems.*

Paraphrased from *Introduction to Clustering Large and High-Dimensional Data* by Jacob Kogan, p. xiv, (2007).

Corollary

Corollary. *Those who believe, or claim, that their method is the best suffer from ignorance and/or arrogance*

Paraphrased from *Introduction to Clustering Large and High-Dimensional Data* by Jacob Kogan, p. xiv, (2007).

Starting Point

A data set where each *element* is described by a vector of numerical *attributes*. Some examples:

Element	Attributes
Movie	Ratings by Netflix Customers
Netflix Customer	Movie Ratings
Cancer Patient	Gene Expression Levels
Iris Flower	Petal and Sepal Measurements
Voting District	Vote Counts for Candidates
Scotch	Flavor Ratings

Notation and a Goal

For this presentation

- A vector data set with n elements each having m attributes will be represented the $m \times n$ matrix $\mathbf{A}$ (i.e. the elements are stored as columns in matrix $\mathbf{A}$).
- Our goal will be to transform $\mathbf{A}$ into an $n \times n$ matrix $\mathbf{S}$ where $\mathbf{S}_{ij}$ gives some measure of the similarity between element $\mathbf{A}_{*i}$ and element $\mathbf{A}_{*j}$.

Measures of similarity between two vectors

- Euclidean distance
- 1-norm
- ∞-norm
- Cosine measure
- Gabriel graph
- A measure derived from a consensus matrix

- Other ideas: Delaunay triangulation, Hamming distance or variation, a new measure you develop

For Comparison's Sake

To make comparison of two similarity matrices easier, I will use the convention that all similarity values lie in the interval $[0, 1]$.

- Cosine and Gabriel measures meet this criteria.
- For a norm, I will use the function $e^{-\frac{||\mathbf{A}_{*i}-\mathbf{A}_{*j}||^2}{2\sigma^2}}$.
  - This is a standard approach to this problem.
  - The σ is a parameter, which can be varied depending on your desire to "spread" similarity values across $[0, 1]$.
  - In this presentation σ is the sample standard deviation.

Why do this?

- Conceptually, you like thinking of the data in a "graph-like" way.

Why do this?

- Conceptually, you like thinking of the data in a "graph-like" way.
- You have powerful graph partitioning software at your disposal.

Why do this?

- Conceptually, you like thinking of the data in a "graph-like" way.
- You have powerful graph partitioning software at your disposal.
- You are a fan of spectral clustering.

Why do this?

- Conceptually, you like thinking of the data in a "graph-like" way.
- You have powerful graph partitioning software at your disposal.
- You are a fan of spectral clustering.
- You need a similarity matrix for the new clustering method you're developing for your Ph.D. thesis.

Running Example

$$\mathbf{A} = \begin{pmatrix} 3562 & 3034 & 2992 & 2730 & 2503 & 2499 \\ 2165 & 2246 & 2062 & 1859 & 2174 & 1276 \\ 4256 & 4189 & 3283 & 2876 & 2873 & 2356 \\ 746 & 724 & 523 & 488 & 506 & 421 \\ 135 & 295 & 140 & 72 & 136 & 47 \\ 160 & 117 & 660 & 511 & 714 & 376 \\ 1314 & 1938 & 1903 & 1860 & 2213 & 1330 \\ 198 & 897 & 338 & 89 & 123 & 128 \\ 1566 & 1249 & 1464 & 1708 & 2062 & 849 \end{pmatrix}$$

Running Example

$$\mathbf{A} = \begin{array}{c} \\ \text{G} \\ \text{R} \\ \text{H} \\ \text{2B} \\ \text{3B} \\ \text{HR} \\ \text{RBI} \\ \text{SB} \\ \text{BB} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{Rose} & \text{Cobb} & \text{Mays} & \text{Ott} & \text{Ruth} & \text{Fisk} \end{array} \\ \begin{pmatrix} 3562 & 3034 & 2992 & 2730 & 2503 & 2499 \\ 2165 & 2246 & 2062 & 1859 & 2174 & 1276 \\ 4256 & 4189 & 3283 & 2876 & 2873 & 2356 \\ 746 & 724 & 523 & 488 & 506 & 421 \\ 135 & 295 & 140 & 72 & 136 & 47 \\ 160 & 117 & 660 & 511 & 714 & 376 \\ 1314 & 1938 & 1903 & 1860 & 2213 & 1330 \\ 198 & 897 & 338 & 89 & 123 & 128 \\ 1566 & 1249 & 1464 & 1708 & 2062 & 849 \end{pmatrix} \end{array}$$

Elements - baseball players.

Attributes - lifetime totals for nine statistics.

A Sample Similarity Matrix

$$\mathbf{S}_{ij} = e^{-\frac{||\mathbf{A}_{*i} - \mathbf{A}_{*j}||_2^2}{2\sigma^2}}$$

	Rose	Cobb	Mays	Ott	Ruth	Fisk
Rose	1.0000	0.4294	0.2787	0.1229	0.0546	0.0174
Cobb	0.4294	1.0000	0.3579	0.1328	0.0848	0.0209
Mays	0.2787	0.3579	1.0000	0.7570	0.5478	0.1871
Ott	0.1229	0.1328	0.7570	1.0000	0.7469	0.3285
Ruth	0.0546	0.0848	0.5478	0.7469	1.0000	0.1051
Fisk	0.0174	0.0209	0.1871	0.3285	0.1051	1.0000

A Sample Similarity Matrix

$$\mathbf{S}_{ij} = e^{-\frac{||\mathbf{A}_{*i} - \mathbf{A}_{*j}||_1^2}{2\sigma^2}}$$

	Rose	Cobb	Mays	Ott	Ruth	Fisk
Rose	1.0000	0.4056	0.2380	0.1084	0.0447	0.0202
Cobb	0.4056	1.0000	0.3239	0.0848	0.0424	0.0074
Mays	0.2380	0.3239	1.0000	0.6804	0.5056	0.0973
Ott	0.1084	0.0848	0.6804	1.0000	0.7083	0.2869
Ruth	0.0447	0.0424	0.5056	0.7083	1.0000	0.1031
Fisk	0.0202	0.0074	0.0973	0.2869	0.1031	1.0000

A Sample Similarity Matrix

$$\mathbf{S}_{ij} = \frac{\mathbf{A}_{*i} \cdot \mathbf{A}_{*j}}{||\mathbf{A}_{*i}||_2 ||\mathbf{A}_{*j}||_2}$$

	Rose	Cobb	Mays	Ott	Ruth	Fisk
Rose	1.0000	0.9838	0.9816	0.9743	0.9480	0.9836
Cobb	0.9838	1.0000	0.9837	0.9703	0.9533	0.9762
Mays	0.9816	0.9837	1.0000	0.9956	0.9848	0.9928
Ott	0.9743	0.9703	0.9956	1.0000	0.9935	0.9866
Ruth	0.9480	0.9533	0.9848	0.9935	1.0000	0.9639
Fisk	0.9836	0.9762	0.9928	0.9866	0.9639	1.0000

A Sample Similarity Matrix

Gabriel graph

	Rose	Cobb	Mays	Ott	Ruth	Fisk
Rose	0	1	1	0	0	0
Cobb	1	0	1	0	0	0
Mays	1	1	0	1	0	0
Ott	0	0	1	0	1	1
Ruth	0	0	0	1	0	0
Fisk	0	0	0	1	0	0

A Sample Similarity Matrix

Consensus matrix built from 100 runs of the NMF algorithm

	Rose	Cobb	Mays	Ott	Ruth	Fisk
Rose	1.0000	0.9800	0.3100	0.1300	0.0300	0.7300
Cobb	0.9800	1.0000	0.3300	0.1500	0.0500	0.7500
Mays	0.3100	0.3300	1.0000	0.8200	0.7200	0.5800
Ott	0.1300	0.1500	0.8200	1.0000	0.9000	0.4000
Ruth	0.0300	0.0500	0.7200	0.9000	1.0000	0.3000
Fisk	0.7300	0.7500	0.5800	0.4000	0.3000	1.0000

The Fiedler Method

- A spectral method

The Fiedler Method

- A spectral method
- Operates on the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{S}$
  - $\mathbf{D} = diag(\mathbf{Se})$
  - $\mathbf{L}$ has row sums of zero

The Fiedler Method

- A spectral method
- Operates on the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{S}$
  - $\mathbf{D} = diag(\mathbf{Se})$
  - $\mathbf{L}$ has row sums of zero
- $\mathbf{L}$ is symmetric, positive semi-definite $\rightarrow$ its eigenvalues are real and non-negative

The Fiedler Method continued

- $\lambda_1 = 0$ and $\lambda_2 > 0$

The Fiedler Method continued

- $\lambda_1 = 0$ and $\lambda_2 > 0$
- Two clusters are formed based on the signs of the entries in $\mathbf{v}_2$, the eigenvector associated with λ_2

$$\mathbf{v}_2 \begin{pmatrix} 0.2548 \\ 0.2034 \\ -0.1414 \\ -0.2093 \\ -0.1905 \\ 0.6077 \end{pmatrix}$$

Fiedler with more than two clusters

- Remove from **S** the rows and columns associated with one cluster. Apply the Fiedler method to the remaining matrix.
- Use the sign patterns of additional eigenvectors to assign clusters. For example, the entries in $\mathbf{v}_2$ and $\mathbf{v}_3$ can be used to find up to four clusters.

$$\begin{array}{cc} \mathbf{v}_2 & \mathbf{v}_3 \\ \begin{pmatrix} 0.5061 & 0.2548 \\ 0.4799 & 0.2034 \\ 0.0687 & -0.1414 \\ -0.2093 & -0.2660 \\ -0.1905 & -0.6585 \\ -0.6548 & 0.6077 \end{pmatrix} \end{array}$$

Clustering the Running Example

Results from applying the Fiedler method to seven different similarity matrices

- {Rose, Cobb, Mays}, {Ott, Ruth, Fisk} - clustering when 1-norm, 2-norm, and Gabriel graph similarity matrices are used
- {Rose, Cobb}, {Mays, Ott, Ruth, Fisk} -clustering when ∞-norm and k-means consensus similarity matrices are used
- {Rose, Cobb, Fisk}, {Mays, Ott, Ruth} - clustering when cosine similarity and NMF consensus matrices are used

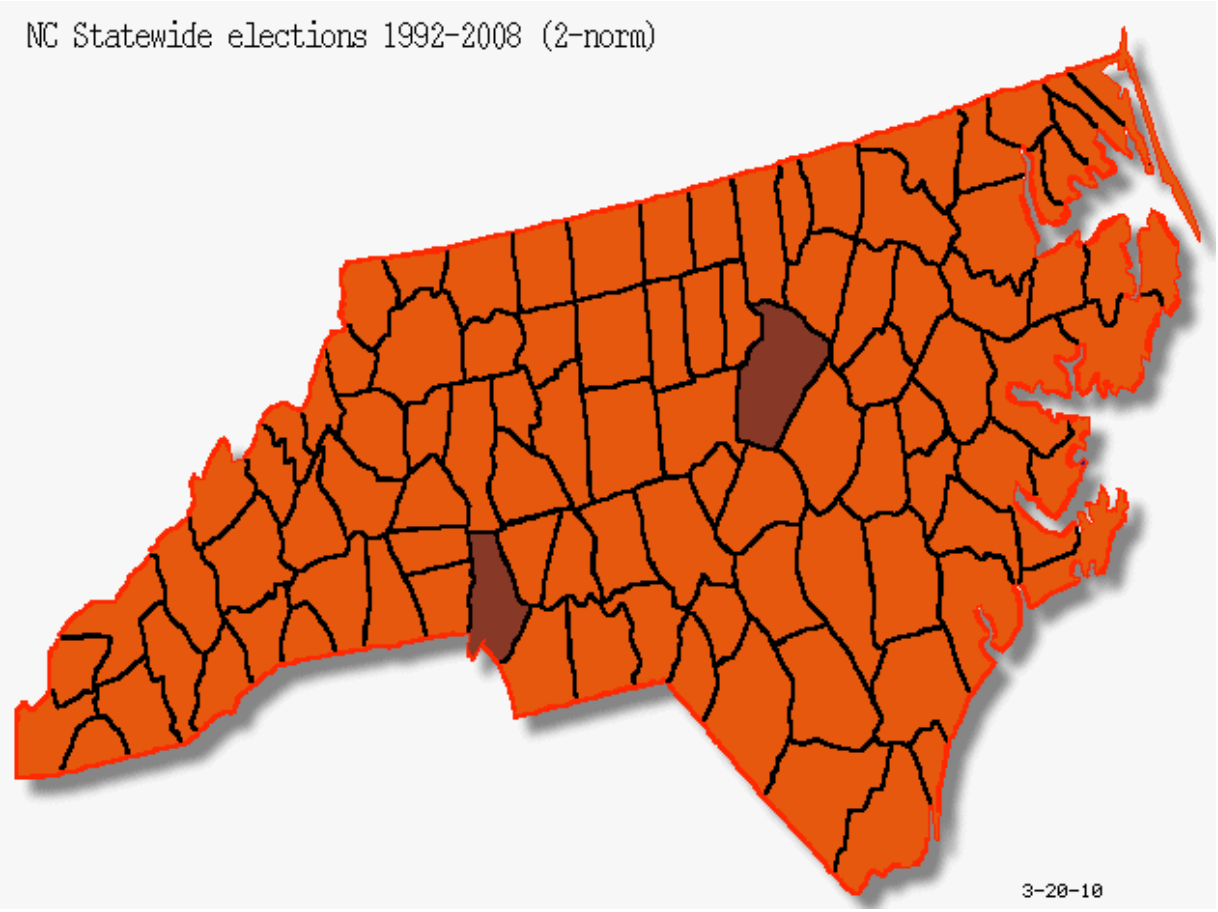
A North Carolina election data set

- All state-wide elections in North Carolina from 1992 to 2008
- Raw vote totals for each candidate in each county
- 16 races (5 presidential, 5 gubernatorial, and 6 U.S. Senate)
- 66 candidates
- Source: North Carolina Department of Elections (http://www.sboe.state.nc.us/content.aspx?id=69)

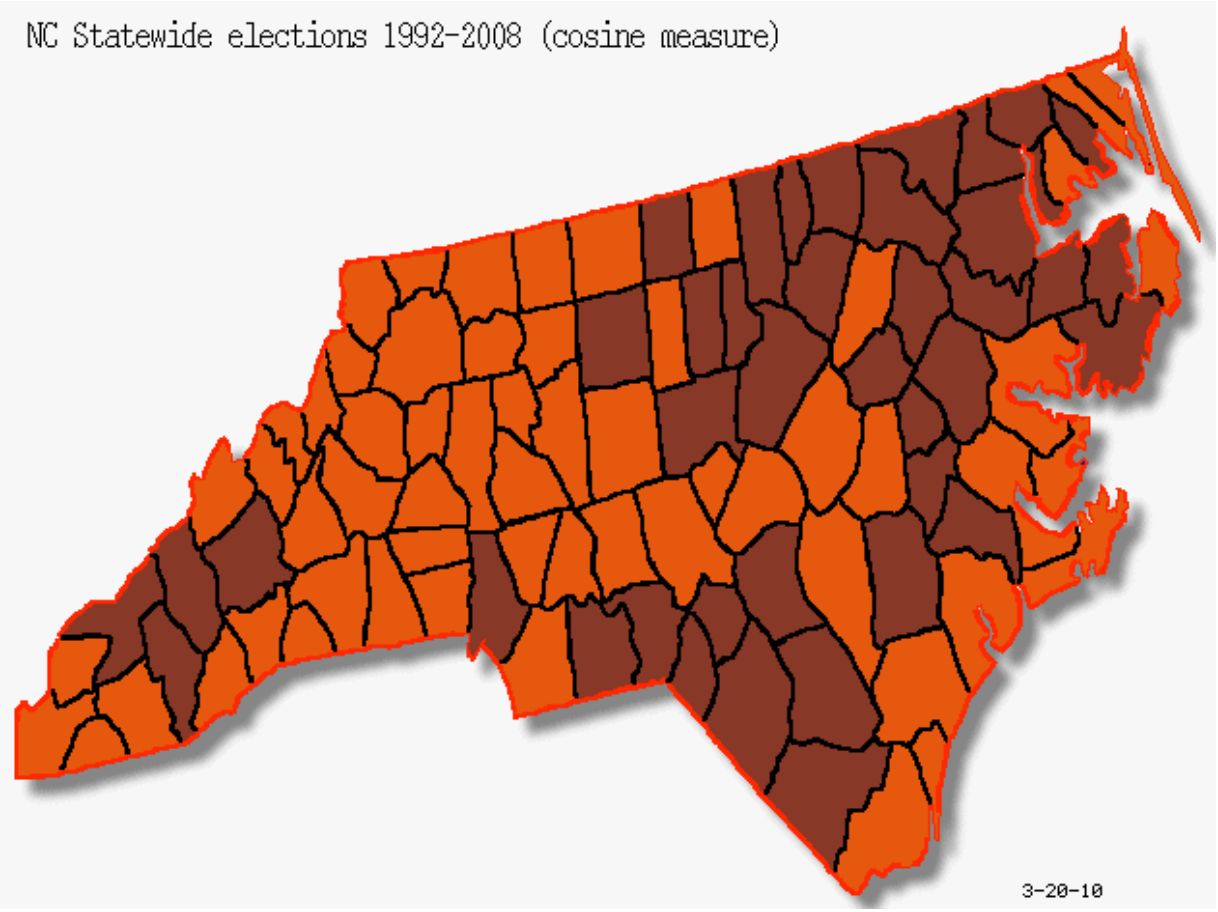
Results

- Goal: Cluster counties with similar voting behavior
- Created seven similarity matrices, using the same seven measures used on the Running Example
  - 1-norm, 2-norm, ∞-norm, cosine measure, Gabriel graph, k-means consensus, NMF consensus
- Clustered using one Fiedler vector

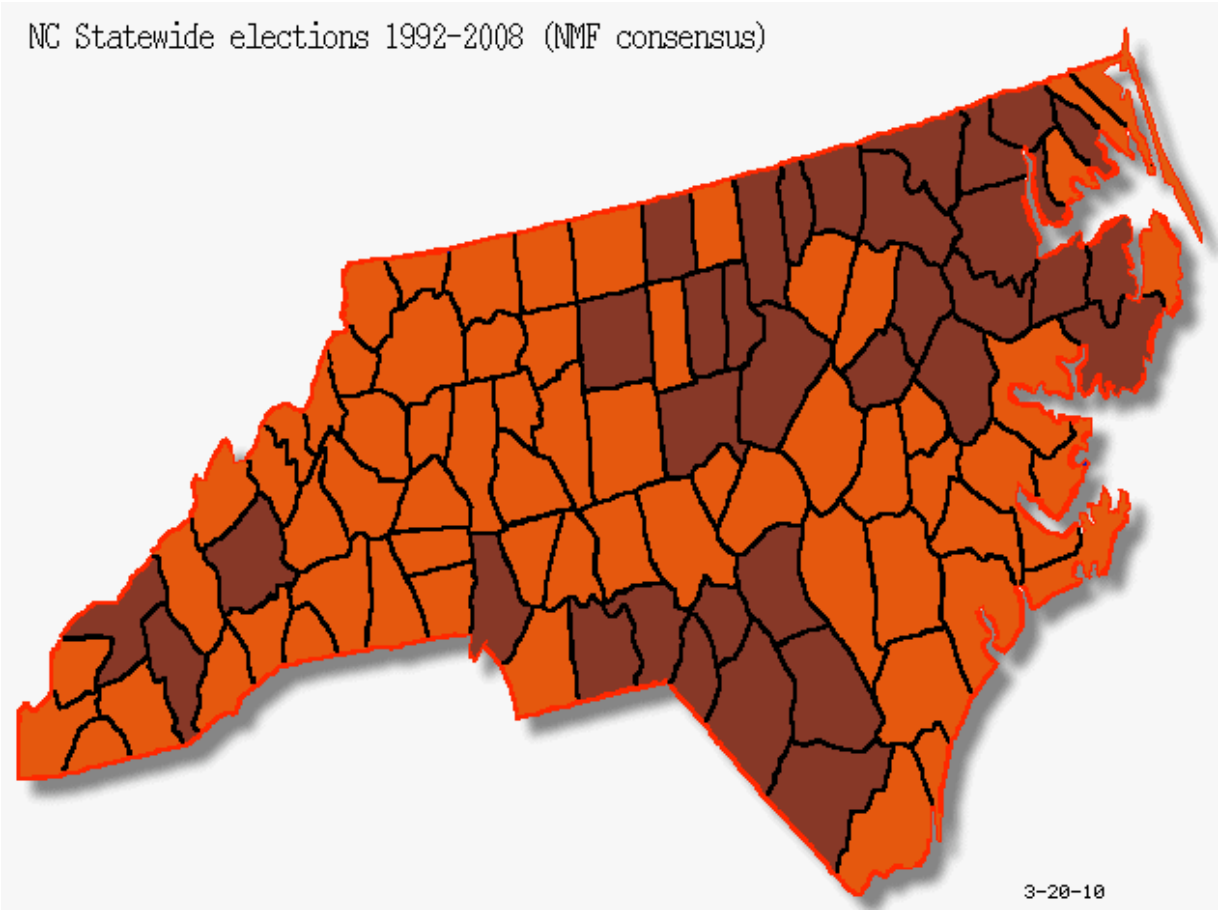
2-norm similarity matrix



Cosine similarity matrix



NMF consensus similarity matrix



The Big Idea, revisited

When clustering a data set with a particular algorithm it is important that

- we have an idea of what we mean by *similarity*.
- the algorithm used meshes with this idea.
- the similarity measure is appropriate for the data being clustered.

Question: Why did using the 2-norm similarity matrix result in such a different clustering?

Conclusion

Lesson for the day:

- Measurements of similarity are not necessarily similar to each other.
- When clustering, consider multiple notions of similarity.
- Consider how your data when choosing a similarity measure.