Clustering Presidential Election Data over Time

Carl D. Meyer
Charles D. Wessell
North Carolina State University

cdwessel@ncsu.edu

MAA SE Section Meeting
March 26, 2010
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
It’s a one-dimensional clustering problem all too familiar to students and professors.

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>101.13</td>
<td>90.72</td>
<td>89.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>88.28</td>
<td>88.28</td>
<td>88.10</td>
<td>85.68</td>
<td>84.38</td>
<td>83.96</td>
<td>80.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>78.34</td>
<td>76.76</td>
<td>75.71</td>
<td>75.13</td>
<td>73.87</td>
<td>71.81</td>
<td>70.57</td>
<td>70.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>66.83</td>
<td>66.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>35.50</td>
<td>23.00</td>
<td>12.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
It’s much more interesting to work with data sets that are large and multi-dimensional. Some examples of such data sets:

<table>
<thead>
<tr>
<th>Element</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>Ratings by Netflix Customers</td>
</tr>
<tr>
<td>Netflix Customer</td>
<td>Movie Ratings</td>
</tr>
<tr>
<td>Cancer Patient</td>
<td>Gene Expression Levels</td>
</tr>
<tr>
<td>Iris Flower</td>
<td>Petal and Sepal Measurements</td>
</tr>
<tr>
<td>Voting District</td>
<td>Vote Counts for Candidates</td>
</tr>
<tr>
<td>Scotch</td>
<td>Flavor Ratings</td>
</tr>
</tbody>
</table>
Some well-known clustering algorithms

1. Hierarchical clustering

2. $k$-means (solutions not unique)

3. Various singular value decomposition (SVD) inspired methods

4. Non-negative Matrix Factorization (NMF) (solutions not unique)
Anyone know what the FToAM is?
Nothing works.
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).
Consensus Clustering

- Use multiple runs of one or more good clustering algorithms to create a better clustering.

- For more background on consensus clustering, see my talk *A Comparison of Consensus Clustering Methods* at http://www4.ncsu.edu/~cdwessel/. (near the bottom of the page)
A presidential election data set

- All U.S. presidential elections from 1912 to 2008
- Vote percentage for each candidate in each of the 48 contiguous states
- 88 total "candidates"
- Source: Dave Leip’s Atlas of U.S. Presidential Elections (http://www.uselectionatlas.org/)
Clustered the states based on each set of five consecutive elections (i.e. 1912-1928, then 1916-1932, etc.)

First created a consensus matrix by running the NMF algorithm 100 times.

Clustered the consensus matrix for the final clustering.
Prelude to the Results

What do you expect to see?
Presidential Election Clustering (1912-1928)
Presidential Election Clustering (1916–1932)

Clustering Presidential Election Data over Time
Presidential Election Clustering (1920–1936)
Clustering Presidential Election Data over Time
Clustering Presidential Election Data over Time
Presidential Election Clustering (1936-1952)
Presidential Election Clustering (1940-1956)
Presidential Election Clustering (1944-1960)
Presidential Election Clustering (1948–1964)
Presidential Election Clustering (1952-1968)
Presidential Election Clustering (1956-1972)
Clustering Presidential Election Data over Time

Presidential Election Clustering (1960-1976)

cdwessel@ncsu.edu, http://www4.ncsu.edu/~cdwessel/
Presidential Election Clustering (1964-1980)
Presidential Election Clustering (1968–1984)

Clustering Presidential Election Data over Time
Presidential Election Clustering (1972-1988)
Clustering Presidential Election Data over Time
Presidential Election Clustering (1980-1996)
1984-2000

Presidential Election Clustering (1984-2000)