A Comparison of Consensus Clustering Methods

Chuck Wessell
Carl Meyer

NCSU

College of Charleston Ranking and Clustering Workshop
August 14, 2009

Outline

- What is consensus clustering?
- Details on the AML_ALL leukemia data set
- A strategy for using consensus clustering on this data set
- Some results with comparisons to other methods

The Big Idea

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

Some Data to Cluster

$$
\begin{array}{l}
\\
\text{Home Runs} \\
\text{RBIs} \\
\text{Average} \\
\text{Stolen Bases} \\
\text{Walks} \\
\text{Games}
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
\text{Ruth} & \text{Cobb} & \text{Mays} & \text{Fisk} & \text{Rose}
\end{array} \\
\begin{pmatrix}
714 & 117 & 660 & 376 & 160 \\
2217 & 1937 & 1903 & 1330 & 1314 \\
.342 & .366 & .302 & .269 & .303 \\
123 & 892 & 338 & 128 & 198 \\
2062 & 1249 & 1464 & 849 & 1566 \\
2503 & 3035 & 2992 & 2499 & 3562
\end{pmatrix}
\end{array}
$$

We could cluster famous baseball players based on their lifetime statistics.

Suppose we cluster this data set and get a clustering of {Ruth, Mays}, {Cobb, Rose} and {Fisk}. We can record this result in an adjacency matrix like this:

$$
\begin{array}{c c}
 & \begin{array}{ccccc} \text{Ruth} & \text{Cobb} & \text{Mays} & \text{Fisk} & \text{Rose} \end{array} \\
\begin{array}{c} \text{Ruth} \\ \text{Cobb} \\ \text{Mays} \\ \text{Fisk} \\ \text{Rose} \end{array} &
\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}
\end{array}
$$

- You can choose to put ones on the main diagonal.
- Original data set: Players defined by their statistics. Adjacency matrix: Players defined by their "proximity" to other players.

After running the a clustering algorithm(s) a number of times, sum all the associated adjacency matrices to obtain the consensus matrix.

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 19 & 36 & 3 & 8 \\ 19 & 0 & 12 & 2 & 41 \\ 36 & 12 & 0 & 8 & 11 \\ 3 & 2 & 8 & 0 & 7 \\ 8 & 41 & 11 & 7 & 0 \end{pmatrix} = C$$

Consensus Clustering

There is not a *consensus* on terminology. A search for 2000-2009 articles using Google Scholar reveals:

- 758 articles containing "ensemble clustering" (203) or "cluster ensemble" (555)
- 662 articles containing "consensus clustering" (468) or "cluster consensus" (194)

For whatever it's worth, Wikipedia has an article entitled "Consensus clustering", but not one entitled "Ensemble clustering".

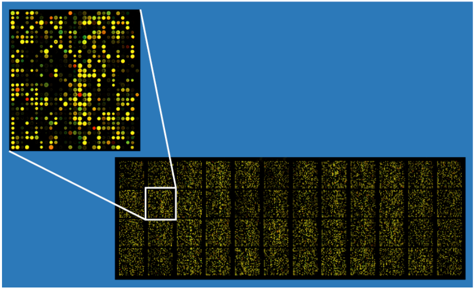
Cluster Aggregation?

For the same time period there were 2480 articles containing "cluster aggregation", but

```
Sci-Tech Dictionary:  cluster aggregation
(physics) A mathematical model of a coagulation process in
which a collection of particles all move randomly at once,
and two particles, or a particle and a previously formed
cluster, stick together whenever they come within a certain
fixed distance of each other.
```

If you search on "cluster aggregation" you have to wade through a lot of physics papers.

A few words on DNA microarrays



The leukemia data set

- First appeared in a 1999 article in *Science*, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring
- Lead authors from MIT's Center for Genome Research which is now part of The Broad Institute, an MIT-Harvard collaboration
- Well known data set in the microarray literature (the 1999 paper is cited in over 6000 other articles)

The leukemia data set

- Contains data from bone marrow samples of 38 cancer patients
- For each sample, gene expression levels for 5000 genes are given (this version of the data set was used in Metagenes and molecular pattern discovery using matrix factorization (2004))
- Non-negative matrix factorization known to do a very good job clustering this data set
- Typical feature of DNA microarray data sets: number of genes is much, much larger than the number of samples

The leukemia data set

The samples can be broken into three groups:

- acute lymphoblastic leukemia, B cell subtype (ALL-B), samples 1-19
- acute lymphoblastic leukemia, T cell subtype (ALL-T), samples 20-27
- acute myelogenous leukemia (AML), samples 28-38

Allows us to evaluate the accuracy of a clustering for either $k = 2$ (ALL/AML) and $k = 3$ (ALL-B/ALL-T/AML).
Note: Sample 29 is a probable misdiagnosis.

The leukemia data set

Note: Sample 29 is a probable misdiagnosis.

Consensus Clustering the leukemia data set

1. Factor the data set using non-negative matrix factorization a “large” number of times with a variety of plausible values of k.
    - Let’s approach this problem as if we do not know the value of k
    - I let k range from 2 to 11 and clustered the data set 10 times for each of these values of k

Consensus Clustering the leukemia data set

2. After each factorization, update the consensus matrix.

Updating the consensus matrix

- After a run of NMF, we examine *H*. One possible result:

$$\begin{array}{cccc} \text{Patient 1} & \text{Patient 2} & \text{Patient 3} & \text{Patient 4} \end{array}$$
$$\begin{pmatrix} .76 & .51 & .21 & .14 \\ .03 & .63 & .81 & .09 \\ .01 & .17 & .09 & .52 \end{pmatrix}$$

- We would update the consensus matrix with the following adjacency matrix:

$$\begin{array}{cc} & \begin{array}{cccc} \text{Patient 1} & \text{Patient 2} & \text{Patient 3} & \text{Patient 4} \end{array} \\ \begin{array}{c} \text{Patient 1} \\ \text{Patient 2} \\ \text{Patient 3} \\ \text{Patient 4} \end{array} & \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Updating the consensus matrix - an alternative

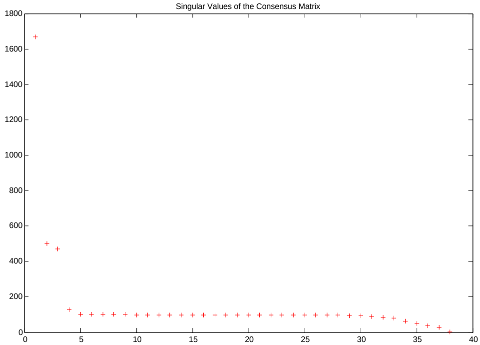
- After a run of NMF, normalize each column of H. One possible result:

$$\begin{array}{cccc} \text{Patient 1} & \text{Patient 2} & \text{Patient 3} & \text{Patient 4} \end{array}$$
$$\begin{pmatrix} .9988 & .6170 & .2452 & .2587 \\ .0457 & .7604 & .9639 & .1657 \\ .0156 & .0721 & .1036 & .9516 \end{pmatrix}$$

- Update the consensus matrix with a matrix of cosines:

$$\begin{array}{c|cccc} & \text{Patient 1} & \text{Patient 2} & \text{Patient 3} & \text{Patient 4} \\ \hline \text{Patient 1} & 0 & .6542 & .2906 & .2808 \\ \text{Patient 2} & .6542 & 0 & .9053 & .4785 \\ \text{Patient 3} & .2906 & .9053 & 0 & .3218 \\ \text{Patient 4} & .2808 & .4785 & .3218 & 0 \end{array}$$

3. Look at the singular values of the consensus matrix to see where their values "level out." Use this to update your interval of plausible *k* values.



Consensus Clustering the leukemia data set

4. If necessary, repeat.
5. Once a single value (or narrow range) for k is settled on, build a final consensus matrix from a “large” number of runs of NMF.
   - For the leukemia data set, this method points to using $k = 3$ clusters. (We will also look at clusterings with $k = 2$ since we know this is a sensible real world clustering.)

6. Cluster your data a "large" number of times, creating a hard consensus matrix along the way and then cluster the consensus matrix.

Some results

Abbreviations used on the following results slides:

- A - the original data set
- CH - the consensus matrix created from hard clusterings
- CC - the consensus matrix created from cosine clusterings

For $k = 2$, a comparison of clustering accuracy between NMF run on the original data, NMF run on two consensus matrices (hard and "cosine") and Fiedler clustering of these same two consensus matrices.

Method	Mean Accuracy	Minimum Accuracy	Maximum Accuracy	Standard Deviation
NMF on A	.9592	.8947	.9737	.0243
NMF on CH	.9111	.5000	.9737	.1264
NMF on CC	.9524	.8947	.9737	.0287
Fiedler on CH	.9737			
Fiedler on CC	.9737			

Some results

For $k = 3$, a comparison of clustering accuracy between NMF run on the original data, NMF run on two consensus matrices (hard and "cosine") and Fiedler clustering of these same two consensus matrices.

Method	Mean Accuracy	Minimum Accuracy	Maximum Accuracy	Standard Deviation
NMF on A	.9282	.6053	.9737	.0791
NMF on CH	.9453	.7368	.9474	.0211
NMF on CC	.9447	.9211	.9737	.0142
Fiedler on CH	.9474			
Fiedler on CC				

Some results

For $k = 4$, a comparison of clustering accuracy between NMF run on the original data, NMF run on two consensus matrices (hard and "cosine") and Fiedler clustering of these same two consensus matrices.

Method	Mean Accuracy	Minimum Accuracy	Maximum Accuracy	Standard Deviation
NMF on A	.8295	.8158	.8947	.0203
NMF on CH	.8163	.8158	.8684	.0053
NMF on CC	.8432	.7632	.9737	.0317
Fiedler on CH	.9211			
Fiedler on CC	.8684			

Some results

Some conclusions based on last few slides.

Other results

If I get something satisfactory, some clustering of North Carolina counties using presidential election data.