K-Means Clustering with Human Tumor Microarray Data

Jill McCracken

September 2009

noblis

For the best of reasons

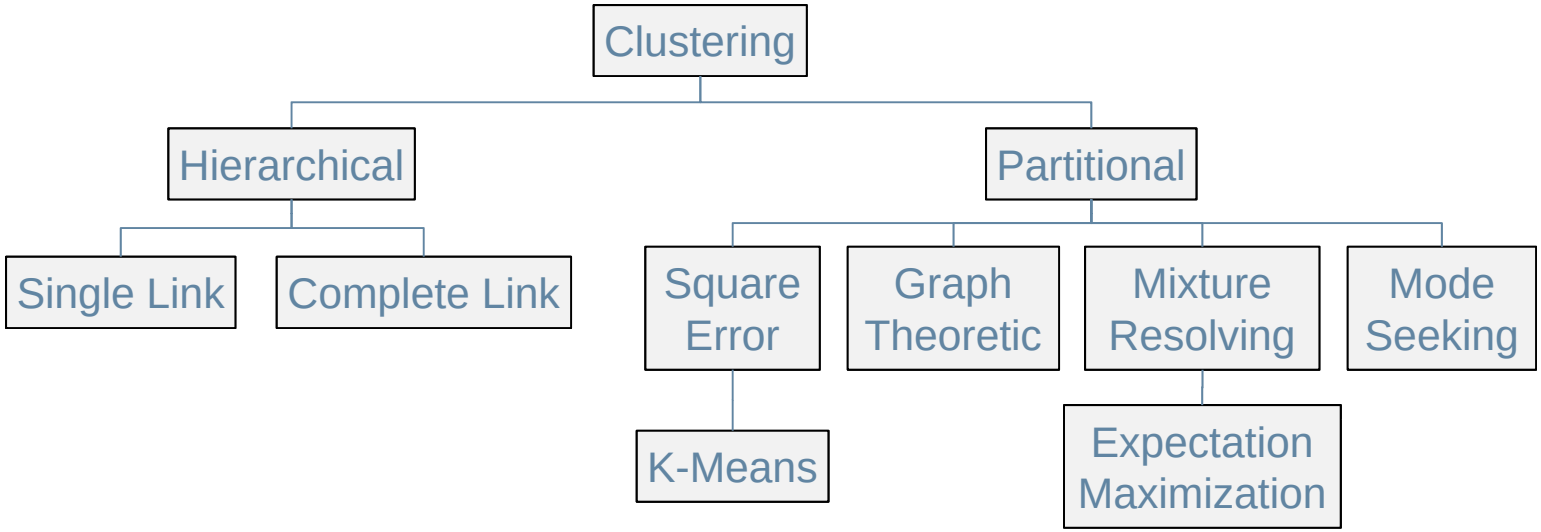
Contents

- K-Means Clustering Overview
- Intro: Human Tumor Microarray Data
- K-Means Clustering Results
- References

Introduction to Clustering

- Objective: assign observations to a cluster such that members of each cluster are more similar to each other than to observations in other clusters (discover natural groupings in the data)
- Analysis Questions:
  - How to assess clustering results?
  - How to prepare the data?
  - Which similarity measure do you use for a given problem?
  - What is the optimal number of clusters?
  - How to apply domain knowledge?
  - How to efficiently cluster large data sets?
  - How to select / extract features?
  - Which clustering algorithm to use?

Jain's Taxonomy of Clustering Approaches [1]



[1] A. K. Jain, M. N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, ***31*** *No. 3, September 1999*

K-Means Clustering

- Parameters specified by the analyst:
  - K = number of clusters
  - Distance measure (Euclidean, Mahalanobis, …)
  - Initial cluster centers
- Basic K-means Algorithm:
  - K seeds are chosen to be the centroids of the clusters
  - Each observation is assigned to a cluster based on which seed is closest
  - New centroids are computed based on cluster assignments
  - Observations are reassigned to clusters based on distances from centroids
  - Stop when squared error stabilizes

K-Means Clustering

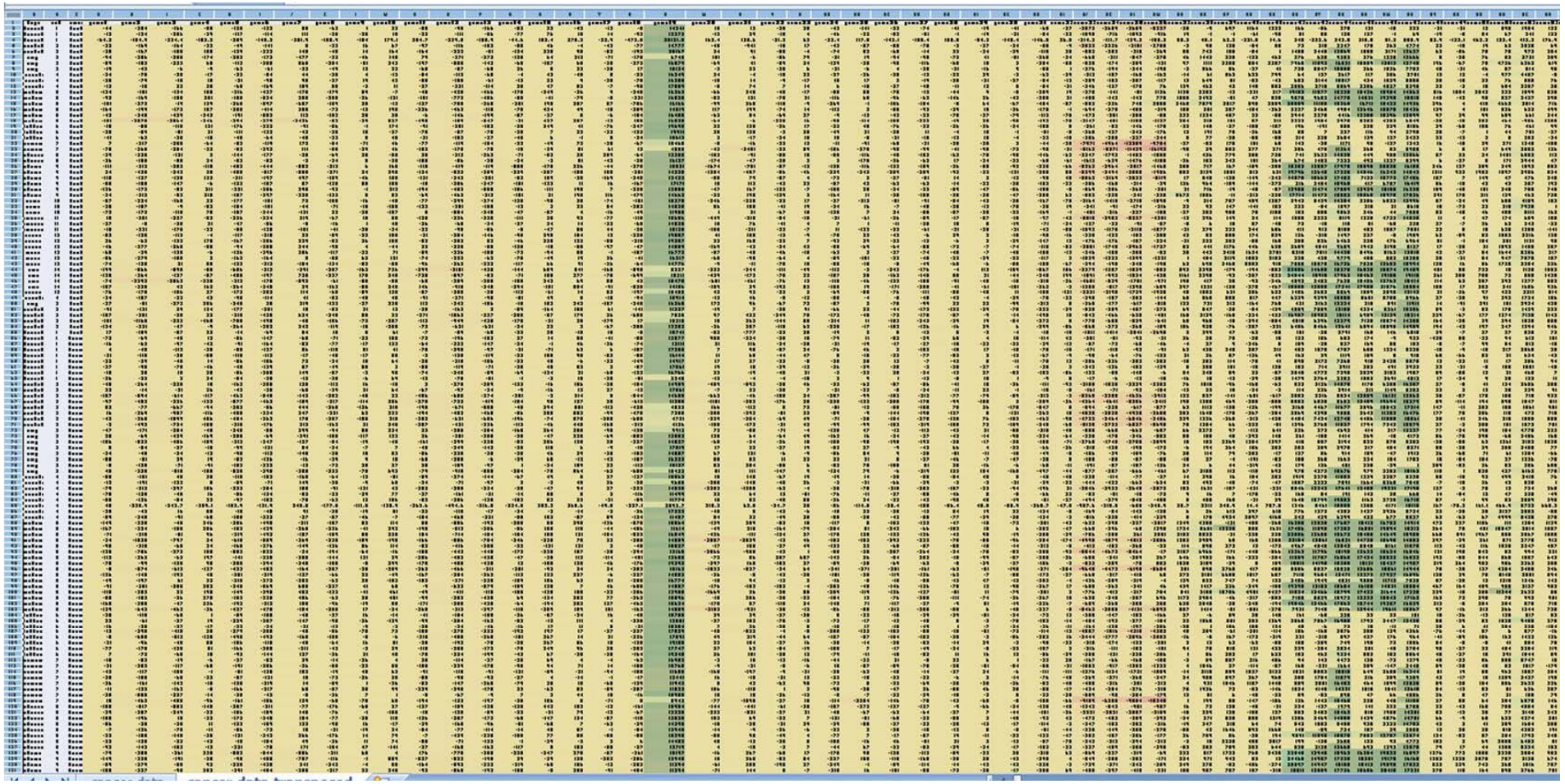
- Pros:
  - Simple algorithm, widely used
  - Relatively fast for large data sets
- Cons:
  - Doesn’t do well with overlapping clusters
  - Produces hyperspherical clusters
  - Results can be sensitive to choice of intial seeds (some variations of k-means aimed at selecting initial seeds that are more likely to lead to global minimum)
  - Clusters are easily pulled off center by outliers

Intro: Human Tumor Microarray Data[3]

- Downloaded from HTF website: http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- DNA microarrays measure the expression of genes in a cell
- 14-cancer gene expression data set:
  - 16064 genes
  - 144 training samples
  - 54 test samples
- One gene per row, one sample per column (transposed prior to analysis)
- Cancer classes are labeled as follows:

1. Breast
2. Prostate
2. Lung
3. Colorectal
4. Lymphoma
5. Bladder
6. Melanoma
8. Uterus
9. Leukemia
10. Renal
11. Pancreas
12. Ovary
13. Meso
14. CNS

Intro: Human Tumor Microarray Data

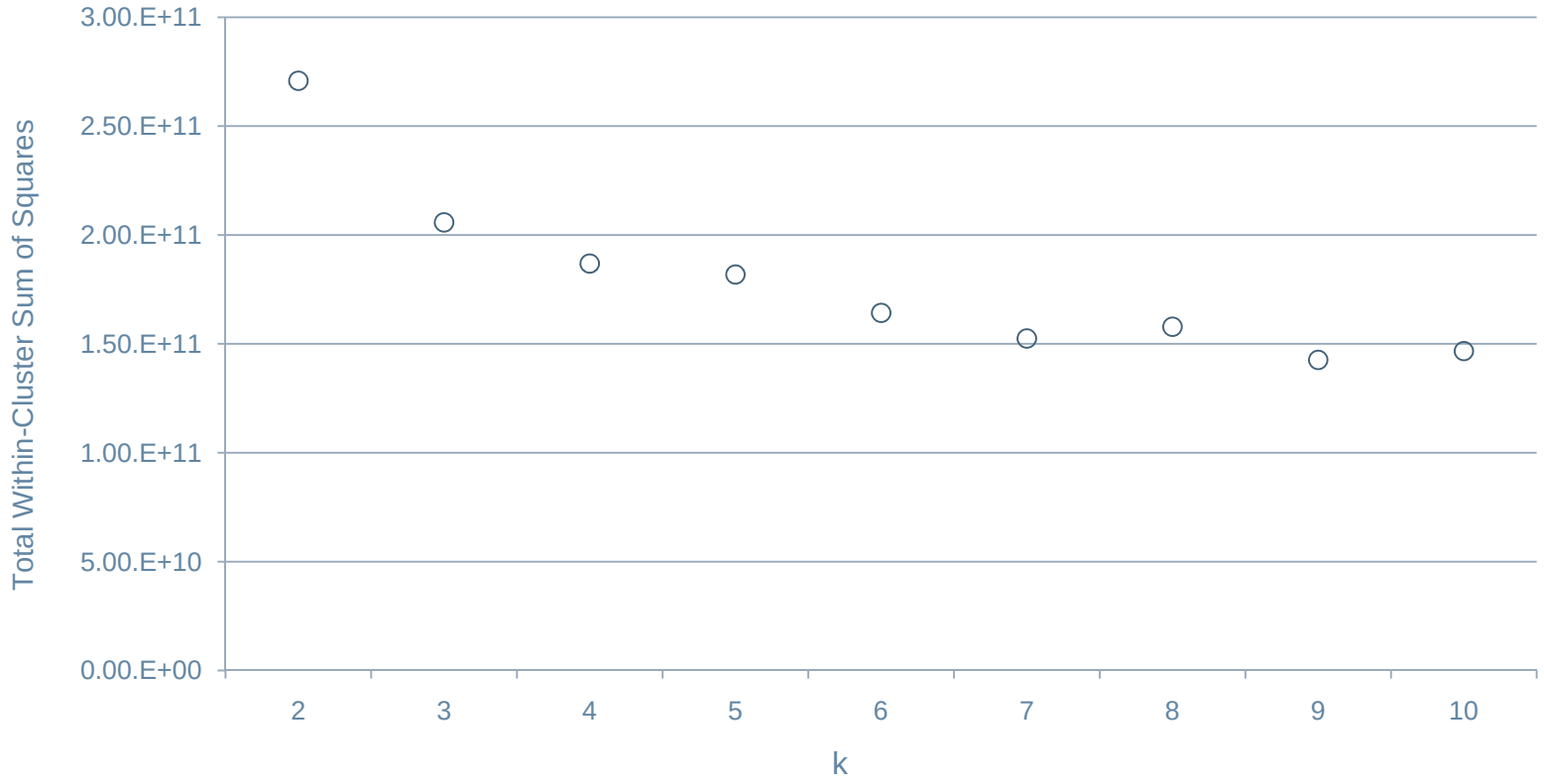


noblis
For the best of reasons

Comparing results for different sets of genes

	Genes 1-5354			Genes 5355-10709			Genes 10710-16064		
	cluster			cluster			cluster		
	Total WithinSS = 3.16 E11			Total WithinSS = 4.16E11			Total WithinSS = 2.06E11		
	1	2	3	1	2	3	1	2	3
bladder	9	2	0	0	0	11	10	0	1
breast	11	0	1	1	0	11	12	0	0
cns	15	5	0	17	0	3	2	8	10
collerectal	5	7	0	3	0	9	9	0	3
lukemia	0	6	24	3	27	0	0	0	30
lung	8	4	0	2	0	10	7	0	5
lymphoma	2	18	2	17	1	4	4	0	18
melanoma	7	3	0	2	0	8	5	0	5
meso	6	5	0	3	0	8	5	0	6
ovary	7	3	2	2	1	9	7	0	5
pancreas	8	3	0	3	0	8	10	0	1
prostate	4	8	2	9	0	5	3	0	11
renal	5	5	1	3	0	8	8	0	3
uterus	4	4	2	3	0	7	7	0	3

Varying k for data set #3



noblis
For the best of reasons

References

1. A. K. Jain, M. N. Murty, and P.J. Flynn, “Data Clustering: A Review,” *ACM Computing Surveys*, **31** No. 3, September 1999

2. A. Jain, “Data Clustering: 50 Years Beyond K-Means,” to be published in *Pattern Recognition Letters*, 2009. url: http://biometrics.cse.msu.edu/JainDataClusteringPRL09.pdf

3. T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning,” Springer, 2009.

4. C. A. Sugar and G. M. James, “Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach,” *Journal of the American Statistical Association* **98** No. 463, September 2003, Theory and Methods.

5. R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” J. R. Statist. Soc. B (2001) **63**, Part 2, pp. 411-423.

6. S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio , W. Gerald, M. Loda, E.S. Lander, and T.R. Golub, “Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures,” *Proc. Natl. Acad. Sci.,* **98**, pp. 15149-15154, 2001.