Beginning a Clustering Literature Survey and a Bisection Clustering Algorithm based on Iterative Correlations

Christopher D. Bell II
Advisor: Dr. Byron Gao

Abstract

As a society, we generate a significant amount of data, but the data must be analyzed for it to be useful. When working with data, there are two cases to consider, where the given data set is labeled and where it is not. Labeling data is an expensive and time-consuming process that requires human effort. To overcome this, data analysis on unlabeled data, or unsupervised learning, can be performed to attempt to find underlying properties of the data.

Clustering is the unsupervised learning method which attempts to partition the data set so that instances within a data set are as similar as possible to each other, and instances across partitions are as dissimilar as possible. There are many clustering methods which can be grouped based on their results.

I have used this rotation project to begin a literature survey on clustering so that I can research the topic, and I used the project to examine a novel bisection clustering algorithm.

Clustering Method Categories

The clustering algorithms that I've examined so far can be separated into six categories: Partitioning, Hierarchical, Density-based, Grid-based, Kernel-based, and Graph Theory-based.

Partitioning Methods are the simplest form of clustering methods. The results of these methods only give the cluster assignment of each instance and tend to be spherical.

Hierarchical Methods either start with all of the data in their individual cluster and combine them, or start with all of the data in a single cluster and divide the clusters. These methods return results in the form of a dendrogram, a chart representing the order in which the clusters were combined or separated.

Density-based methods consider how dense the space around each instance is with other instances. These methods can find arbitrary shapes in data, where hierarchical and partitioning methods cannot.

Grid-based methods group the instances into cells and perform analysis on these cells. These methods can handle high-dimensional data sets.

Kernel-based methods use a kernel function to map the dataset to a higher dimensional space with the objective of making the dataset linearly separable.

Graph Theory-based methods examine a graph representation of the data for various graph phenomena, to include highly connected subgraphs or the spectrum of the graph's Laplacian matrix.

Clustering Methods

Partitioning Methods

- k-Means
- k-Medoids
- k-Medians
- CLARA: Clustering Large Applications
- CLARANS: Clustering Large Applications upon Randomized Search
- ISODATA: Iterative self-organizing data analysis techniques

Hierarchical Methods

1. Top-Down (Divisive)
   - DIANA: Divisive Analysis
   - MONA: Monothetic Analysis
2. Bottom-Up (Agglomerative)
   - AGNES: Agglomerative Nesting
   - BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies
   - CURE: Clustering Using Representatives
   - ROCK: Robust Clustering Algorithm for Categorical Attributes

Density-based Methods

- DBSCAN: Density-based Spatial Clustering of Applications with Noise
- OPTICS: Ordering Points to Identify the Clustering Structure
- DENCLUE: Clustering Based on Density Distribution Functions

Grid-based Methods

- STING: Statistical Information Grid
- CLIQUE: An Apriori-like Subspace Clustering Method

Graph-based Methods

- Chameleon
- HCS: Highly Connected Subgraphs
- CLICK: Clustering Identification via Connectivity Kernels
- Spectral Clustering

Kernel-based Methods

- Kernel k-Means
- Support Vector Clustering

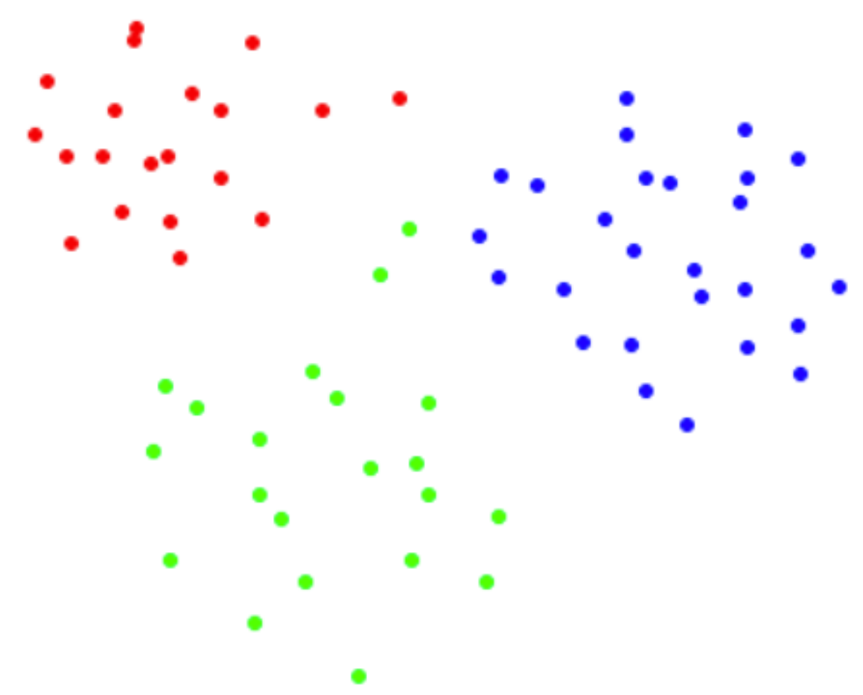

Image 1. Result of k-Means on a synthetic dataset

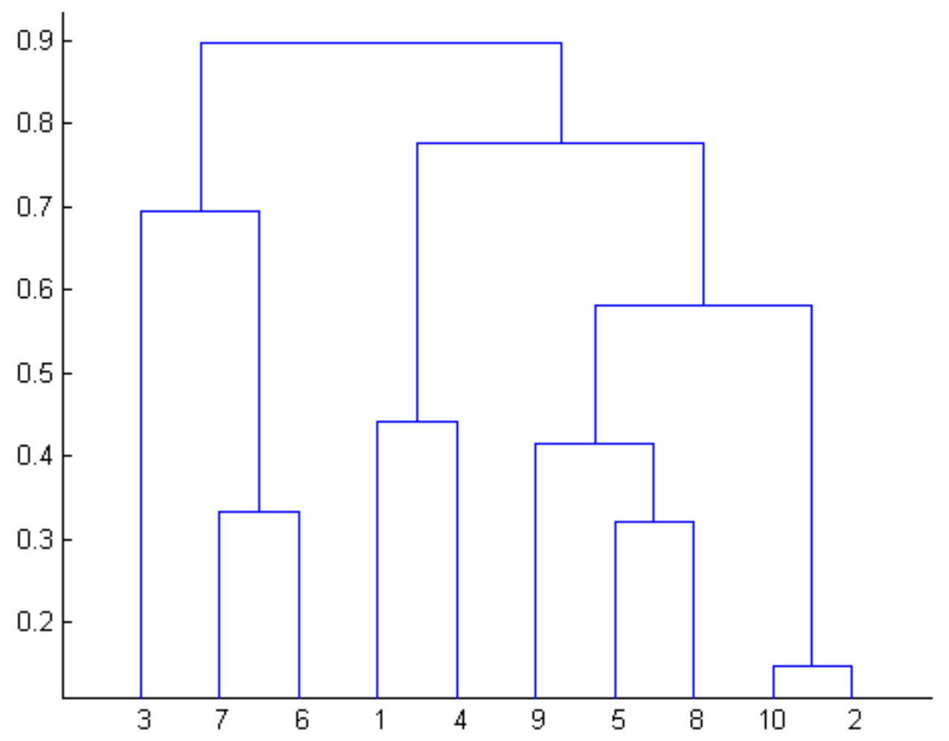

Chart 1. An example dendrogram
Image from: https://i.stack.imgur.com/Cwakg.png

Bisecting Data using Iterative Correlations

In a dataset, if two instances share similar distances between themselves and other instances, then it is likely that the instances themselves are similar. Using this idea, we can calculate the correlation between two instance's rows in a distance matrix to examine how similar they are.

The result of the correlation calculation for each instance pair is a matrix of correlations. Because these correlations also represent a similarity between the instances, we can reapply this concept.

By iteratively performing this correlation calculation and rounding the value to 1 or -1 once we are within some epsilon threshold of either, the resulting correlation matrix converges. We can then use any column of the matrix to assign a label to each instance, bisecting the set.

	1	2	3	4	5	6	7	8
1	0.0	48.8	59.5	35.2	253.0	246.3	271.3	309.9
2	48.8	0.0	58.0	54.1	264.6	250.7	277.1	314.4
3	59.5	58.0	0.0	26.9	206.8	194.4	220.5	258.2
4	35.2	54.1	26.9	0.0	220.2	211.9	237.2	275.6
5	253.0	264.6	206.8	220.2	0.0	39.8	35.9	68.4
6	246.3	250.7	194.4	211.9	39.8	0.0	27.0	63.9
7	271.3	277.1	220.5	237.2	35.9	27.0	0.0	38.9
8	309.9	314.4	258.2	275.6	68.4	63.9	38.9	0.0

Table 1. The distance matrix for the data in figure 2.

Notice the values in rows 1 and 2 are similar (Corr = 0.9779)

	1	2	3	4	5	6	7	8
1	1	1	1	1	-1	-1	-1	-1
2	1	1	1	1	-1	-1	-1	-1
3	1	1	1	1	-1	-1	-1	-1
4	1	1	1	1	-1	-1	-1	-1
5	-1	-1	-1	-1	1	1	1	1
6	-1	-1	-1	-1	1	1	1	1
7	-1	-1	-1	-1	1	1	1	1
8	-1	-1	-1	-1	1	1	1	1

Table 2. The converged iterative correlation matrix of Table 1

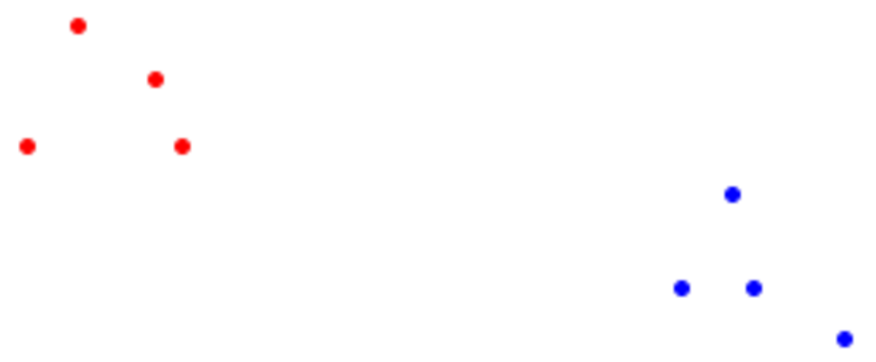

Figure 2. A synthetic dataset with a bisection based upon labels from Table 2

Future Work

- Extend literature review to the rest of the seminal work, and then move onto modern topics in clustering, such as subspace and stream clustering
- Add bisecting clustering algorithm to top-down hierarchical clustering algorithms, and examine results
- Use cluster analysis on documents to partition document search results for user

Contact

Christopher D. Bell II
Texas State University
Email: chris-bell@txstate.edu
Website: cs.txstate.edu/~cdb125/

References

1. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in Proc. ACM SIGMOD Int. Conf. Management of Data, 1998, pp. 94–105.
2. G. Ball and D. Hall, "A clustering technique for summarizing multivariate data," Behav. Sci., vol. 12, pp. 153–155, 1967.
3. M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD'96), 1996, pp. 226–231.
4. S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. ACM SIGMOD Int. Conf. Management of Data, 1998, pp. 73–84
5. S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," Inf. Syst., vol. 25, no. 5, pp. 345–366, 2000.
6. J. Han, M. Kamber, and J. Pei, Data mining: Concepts and Techniques: Elsevier/Morgan Kaufmann, 3rd ed, 2012
7. E. Hartuv and R. Shamir, "A clustering algorithm based on graph connectivity," Inf. Process. Lett., vol. 76, pp. 175–181, 2000.
8. A. Hinneburg and D. Keim, "An efficient approach to clustering in large multimedia databases with noise," in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD'98), 1998, pp. 58–65.
9. G. Karypis, E. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," IEEE Computer, vol. 32, no. 8, pp. 68–75, Aug. 1999.
10. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp., vol. 1, 1967, pp. 281–297.
11. R. Sharan and R. Shamir, "CLICK: A clustering algorithm with applications to gene expression analysis," in Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology, 2000, pp. 307–316.
12. R. Xu and D. Wunsch II, "Survey of Clustering Algorithms," IEEE TNN, vol. 16, no. 3, May 2005.
13. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Conf. Management of Data, 1996, pp. 103–114.