CS410 Text Clustering Review Andrew Chandra

arc11@illinois.edu November 7th 2021

The goal of clustering is to group documents or terms by their similarity. There are two main approaches to clustering: using a similarity function (K-means, Hierarchical Agglomerative Clustering) or based on generative probabilistic models (e.g. Brown, Gaussian mixture model, Neural language model). Each approach is chosen based on the nature of the problem, with models generally requiring more computational resources. Effective clustering can be measured by how compact/dense the clusters are (intra-cluster variance) and how well-separated the clusters are. Objects in a cluster should be highly similar and the overall clustering should provide us insight or utility.

K-means hard circular clustering aims to partition our data into k clusters based on similarity and mutual exclusion from the other clusters. Each cluster is formed around a centroid value (the mean of the data) and observations are sorted using a modified unsupervised K-nearest neighbors (KNN) or greedy heuristic algorithm which sorts based on a robust distance (e.g. Manhattan) from the centroid. The analyst must select a good value for the number of clusters, k, maximizing cluster similarity without overfitting. K-means is also sensitive to outliers.

An alternative approach to partitioning is the DBSCAN method which will form clusters based on their density: points are designated as core, non-core and outliers based on their radial distance and reachability from a core point originating a cluster. Only core points can reach a non-core point, forming a directed acyclic graph. All points in the cluster are reachable by the originating core point. A major advantage of DBSCAN compared to K-means is that we don’t have to prespecify the number of clusters.

Hierarchical clustering of data into dendrograms can be approached in two ways: top-down, starting from one cluster and creating many as with DIANA and MONA (for binary variables); bottom up or agglomerative (merging clusters) seen with BIRCH and CURE. BIRCH (for numeric data) and CURE are a major improvement over AGNES, providing a faster time complexity, allowing their use on larger datasets. Brown clustering is a hierarchical method that will cluster a *vocabulary* into word classes/clusters, forming correlations between words and identifying context.

Gaussian mixture modeling (GMM) assumes a normal curve for the data points and uses the Expectation-Maximization local optimization algorithm to find the mean and variance and iteratively associate data points with soft clusters that are potentially non-spherical and overlapping based on the Gaussian model.

Selecting the appropriate clustering method depends on the constraints of the problem, the computational resources available, and the nature of the data (website, document, sentence, word). K-means is a good starting point that may soon prove inadequate, leading us to more complex methods that risk overfitting and are more difficult to interpret. Effective clustering will provide structure to the internet, enabling new ways to access and explore information; power recommendation engines; make sense of market data; and improve medical treatments and bioinformatic applications.

References

Ester, M., Kriegel, H., Sander, J. and Xu, X., 1996. *A density-based algorithm for discovering clusters in large spatial databases with noise*. [online] Www2.cs.uh.edu. Available at: <http://www2.cs.uh.edu/~ceick/7363/Papers/dbscan.pdf> [Accessed 7 November 2021].

Zhang, T. and Ramakrishnan, R., 2021. *BIRCH: An Efficient Data Clustering Method for Very Large Databases*. [online] Www2.cs.sfu.ca. Available at: <https://www2.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf> [Accessed 7 November 2021].

Mackey, L., 2014. *Stats 306B: Methods for Applied Statistics: Unsupervised Learning; Gaussian Mixture Models; Expectation-Maximization*. [online] Web.stanford.edu. Available at: <https://web.stanford.edu/~lmackey/stats306b/doc/stats306b-spring14-lecture2_scribed.pdf> [Accessed 7 November 2021].