Exploring Clustering Methods

Brief Introduction: Clustering Concept

Clustering analysis as an unsupervised learning has been shown (Provost and Fawcett, 2013) to be an effective method in creating customer segmentation, as it helps extract natural subgroups of similar data objects. Furthermore, in domains with large population of customers, properly dividing the market into analogous segments could help better manage the forecasts generated and improve the prediction accuracy (Murray, Agard and Barajas, 2018). Despite a variety of clustering algorithms to choose from, two of the most widely used techniques are K-means Clustering and Hierarchical Clustering (Tripathi, Bhardwaj and Poovammal, 2018).

K-means Clustering

K-means Clustering is a type of centroid-based clustering. The mechanism (Provost and Fawcett, 2013) starts from creating k initial cluster centers. Next, for each of the clusters, its cluster center, or centroid, is calculated, activating the assignment of the data points to their belonging group. The process of classification and centroid adjustment simply iterates, until no change is made in the clusters or some stopping criteria is met. The desirable result of clusters should achieve the objective that, "the inter-group homogeneity is maximized and the intra-group heterogeneity is also maximized (Murray, Agard and Barajas, 2018)."

Hierarchical Clustering

Hierarchical Clustering tries to form a hierarchy of clusters (dendrogram) based on similarities found in data points, in the sense that it considers not only the similarities between the individual instances, but also the linkages between the individual clusters. Such algorithms can be broadly classified into two categories (Cohen-addad, Kanade, Mallmann-trenn and Mathieu, 2019), "agglomerative approaches which grow the cluster tree bottom-up, and divisive approaches which grow the cluster tree top-down."

Data Type: Categorical Data

For numeric data, K-means Clustering is often the efficient approach when one is tasked with processing large datasets. For categorical data, the selection of appropriate algorithm requires special consideration since there does not exist the obvious approach. Still, the three main solutions have been proposed and evaluated against one another.

K-means Clustering (Binary-based)

To preserve its nice feature of efficiency, researchers (Ralambondrainy, 1995) decide to continue to implement K-means algorithm, which requires manipulating data in the way that each categorical attribute is translated into several binary attributes. The prescribed dummification procedure will then allow the transformed data to be fed into either the K-means Clustering (Dissimilarity measure: Euclidean distance) or the Hierarchical Clustering (Dissimilarity measure: Jaccard index).

K-modes Clustering

Though the aforementioned method could produce results, it still suffers from some problems and limitations.

Hierarchical Clustering

Even though the algorithm designed can already handle data with numeric and categorical values, it comes with the quadratic computational cost (Badase, Deshbhratar and Bhagat, 2015). As a result, it might not be an ideal choice when one has to perform the clustering analysis on a large dataset. Furthermore, the lack of a well-defined mathematical objective function might impede the overall model evaluation (Cohen-addad, Kanade, Mallmann-trenn and Mathieu, 2019).

K-means Clustering

First, converting multiple categorical attributes to binary attributes might result in a large number of binary attributes, an undesirable feature that again increases the computational and space costs. Second, due to the dummification, the cluster means do not carry the clear meaning, and thus no longer indicate the characteristics of the clusters.

Given the above pitfalls, researchers develop a new algorithm, K-modes Clustering (Huang, 1998), making some modifications to the K-means Clustering.

- Using a simple matching criterion to measure the distance between categorical data objects
- Replacing means of cluster with modes of cluster
- Using a frequency-based method to find the modes and update the centroids

K-means Clustering (Relative-frequency-based)

Like K-means Clustering, K-modes Clustering is not without its own limitations (Salem, Naouali and Sallami, 2017). For one, the resultant clusters are only local optimal, which does not cover the global information thoroughly. For another, the obtained results are still sensitive to the number of and the shape of initial centroids chosen. Besides, the simple matching dissimilarity measure adopted in the K-modes algorithm often results in clusters with weak intra-similarity (Ng et al., 2007). To accommodate the last point raised, researchers (Salem, Naouali and Sallami, 2017) turn back to the K-means algorithm, with each categorical value being encoded with its corresponding frequency. Compared to the K-means Clustering with pure dummification, K-means Clustering with relative-frequency encoding achieves higher accuracy and less complexity.

Evaluation Criteria

Due to the lack of external class labels, internal cluster validation methods are applied to help determine the appropriate/optimal number of clusters. Two of the most common ones are elbow method and silhouette width, with the former focusing on the cluster-level performance, and the latter centering on the observation-level performance.

Elbow method

The statistics of interest is the total within cluster sum of squares, which calculates the total distances between each data point and its cluster center (centroid) across clusters and thus measures the compactness/goodness of the clustering (Boehmke, 2016). As the number of clusters (k) increases, data points in their respective clusters are expected to connect closely to each other, resulting in a smaller value of total within cluster sum of squares. To help determine the optimal number of clusters, one can create the elbow plot, which plots the statistics against the multiple values of k, and find the point at which the curve starts to flatten out.

Average silhouette method

The silhouette width (Rousseeuw, 1987) serves as an additional evaluation method. Taking both the average within cluster distance and the average closest neighbor distance into consideration, the statistic tries to measure how well each observation fits into its assigned cluster. For each observation, the silhouette width (S(i)) is determined as follows:

$$1 - (C(i) / N(i)), \text{ if } C(i) < N(i)$$

$$0 \qquad\qquad , \text{ if } C(i) = N(i)$$

$$(N(i) / C(i)) - 1, \text{ if } C(i) > N(i)$$

or written in one formula,

$$S(i) = N(i) - C(i) / \max\{C(i), N(i)\}$$

where C(i) denotes the average dissimilarity between the observation *i* and all other observations in the same cluster, and N(i) denotes the average dissimilarity between the observation *i* and all observations in the neighboring cluster. The interpretation of silhouette width is that:

- S(i) is close to 1: The observation is well assigned to its current cluster (“well-clustered”).
- S(i) is close to 0: The observation is on the borderline between two clusters, and so it is unclear whether it should have been assigned to which cluster (“intermediate case”).
- S(i) is close to -1: The observation is mis-assigned to its current cluster, and therefore should be assigned to the neighboring cluster (“mis-clustered”).

Bibliography

Badase, P.S., Deshbhratar, G.P. and Bhagat, A.P., 2015, March. Classification and analysis of clustering algorithms for large datasets. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (pp. 1-5). IEEE.

Boehmke, B., 2016. *K-Means Cluster Analysis · UC Business Analytics R Programming Guide*. [online] Uc-r.github.io. Available at: https://uc-r.github.io/kmeans_clustering [Accessed 10 May 2020].

Cohen-addad, V., Kanade, V., Mallmann-trenn, F. and Mathieu, C., 2019. Hierarchical Clustering. *Journal of the ACM*, 66(4), pp.1-42.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), pp.283-304.

Murray, P.W., Agard, B. and Barajas, M.A., 2018. Forecast of individual customer's demand from a large and noisy dataset. *Computers & Industrial Engineering*, *118*, pp.33-43.

Ng, M.K., Li, M.J., Huang, J.Z. and He, Z., 2007. On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), pp.503-507.

Provost, F. and Fawcett, T., 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.".

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, pp.53-65.

Ralambondrainy, H., 1995. A conceptual version of the K-means algorithm. *Pattern Recognition Letters*, *16*(11), pp.1147-1157.

Salem, S.B., Naouali, S. and Sallami, M., 2017. Clustering categorical data using the k-means algorithm and the attribute's relative frequency. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *11*(6), pp.708-713.

Tripathi, S., Bhardwaj, A. and Poovammal, E., 2018. Approaches to clustering in customer segmentation. *International Journal of Engineering & Technology*, *7*(3.12), pp.802-807.