Reasons for overlapping

Overlapping clusters: This implies that items will not belong to one particular cluster but to multiple clusters, so we can say that when ever these items belongs to multiple clusters. We talk about overlapping when two clusters are present in the same area of the data space

Several reasons can produce overlapping clusters:

* there might be noise in the data,
* the features may not capture all the necessary information to clearly separate clusters
* the overlap may be inherent to the processes that produced the data.

Overlapping is a problem for algorithms that assume a clear separation of the clusters, or at least a zone of lower density points.

overlapping clusters, as shown in figure 2. Overlapping clusters is an issue that is not often tackled by clustering algorithms. Some state-of-the-art algorithms such as spectral clustering [18], which is very good at discovering arbitrary shaped clusters, will fail in the presence of overlapping. On the contrary, the EM [6] algorithm has a bias towards spherical clusters but can handle overlapping quite well as we show in section 3.

3 Metrics evaluation, what is the highest attained score from each of the metrics used

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | KMeans | DBSCAN | Agglomerative |
| Silhouette\_score | 0.20959 | 0.18340 | 0.19392 |
| Davies Bouldin\_score | 1.912 | 1.809 | 1.966 |
| Calinski Harabasz\_score | 1706.148 | 808.150 | 1568.766 |

From the above result gotting from the metric evaluation we see that Kmeans is said to be the best clustering method when compared to the other two techniques. We have the silhouette score to be the highest for the KMeans when compared to the other individual silhouette score, and followed by the Calinskki Harabasz score having a value of 1706.148 higher than that of DBSCAN (808.150) and aslo the Agglomerative (1568.766). With DBSCAN having the lowest performance level judging from all metrics used for evaluation. A silhouette score of 0.18340 , Davis Bouldin Score of 1.809 and Calinski Harabasz Score of 808.150.

Explanation of cluster visualization

What the clusters are based on (customer behavior or buying pattern)

The clustering is based on checking the purchasing pattern of various customers. Patterns that has to do with customers installmental payments, purchase frequency, payment methods and the like as seen in the data set.

What this clustering is helping business owners do, what decision can be made from the algorithm, what story can be told

Knowledge contribution

Recommendation for further research

For future research some of the additional effects that can be performed on this analysis are as follows:

* Trying out other feature selections techniques
* Using other scaling techniques in scaling the data.
* More hyper parameter tuning.
* Involving techniques to handle and treate overlapping such as the EM clustering implementing.

Analysis conclusion

* Using the Silhouette evaluation metric the KMeans is said to have perform better than both the DBSCAN and Agglomerative with values of 0.18 and 0.19, while for means we have a slightly higher increase in performance of 0.209
* The DBSCAN is seen to perform better than the other two clustering algorithms with a lowest value in its **Davies Bouldin Index** of 1.8. The **Davies Bouldin Index implies that the lower the value the better the performance.**
* **The Calinski implies that the higher the value the better the performance so therefore, the KMeans for the second time is seen to be the best clustering technique.**

In conclusion having tested three metric and two of the metrics indicated Kmeans as the best on their various levels, we could conclude with some level of significance that KMeans is seen to be the best clustering technique in this case study.

Visualization of result