CSCI 347: Final Project Report

Tanis Hadwin, Jacob Clostio, Tyler Koon, and Chris Vazquez

Introduction

As datasets continue to grow in size and complexity, the use of traditional supervised learning techniques for large-scale research efforts is becoming increasingly expensive and time consuming. To address these challenges, many projects have begun to look towards unsupervised learning methods which are better equipped for unlabeled and complex data. Among these methods, K-Means clustering has perhaps been the most well-adopted, largely due to its simplicity, flexibility, and performance. However, the default K-Means algorithm is not without its limitations. Most notably, the random initialization of cluster centroids exposes the algorithm to getting caught in local, non-optimal minima. This can severely harm clustering performance, especially with highly-clustered or high-dimensional data (where clusters may be very close together). Our research attempts to better understand the impact of this problem and how it might be addressed by comparing the performance between K-Means and two alternative implementations: K-Means++ and the Hartigan Wong K-Means algorithm.

Dataset and Problem Domain

To explore this research question, we considered the performance of these algorithms in the context of the Original Breast Cancer Wisconsin dataset from the UCI Machine Learning Repository. This dataset consists of 699 instances of tumorous masses collected from patients at the University of Wisconsin Hospitals. Each sample is represented by nine numerical attributes that record the physical and cellular properties of the mass, and a single binary response variable that classifies the sample as either benign or malignant. This access to the ground truth assignment for each sample was highly desirable for the purpose of this project as it allowed for more powerful evaluation metrics to be used. This in-turn provided more accurate comparisons of each method's performance. Additionally, the smaller size of this dataset supported the time constraints imposed on our research efforts while still representing a sufficient sample space to provide meaningful results.

Although the data satisfied the conditions of the project, it still required preprocessing before it could be used in the experiment. The first step in our preprocessing pipeline involved dropping irrelevant attribute data that contributed insignificant information to the underlying analysis. For

these data in particular, a patient identifier attribute was dropped from the dataset to reduce overall complexity during the clustering process. Additionally, the 16 samples that were missing attribute data were also removed from the dataset. Because these samples only comprised around two percent of the total dataset, we believed that removing them would have little impact on our results. With the incomplete samples removed and the irrelevant attributes pruned, we prepared for dimensionality reduction by applying standard normalization. We favored this normalization technique due to its ability to preserve the relationships for variation within and between attributes, which might otherwise be lost when using simpler techniques such as range normalization. The final step in our preprocessing involved reducing the dimensionality of these data to two dimensions. Besides reducing complexity, this allowed for a more digestible analysis and visual representation of the underlying clustering behavior.

Experiment Design

The three methods considered in our experiment were K-Means, K-Means++, and Hartigan Wong K-Means. As previously discussed, K-Means randomly initializes cluster centroids which can lead to suboptimal clustering results due to centroids being trapped in local minima. K-Means++ works by assigning the first cluster centroid at random, then selects centroids based on a probability distribution that is proportional to the squared distance of the nearest existing centroid. That is, centroids will be more likely to be initialized further away from the mean of existing centroids. Lastly, Hartigan Wong K-Means systematically initializes centroids by randomly assigning each point to a cluster, and then defining cluster centroids according to the average position of each cluster member. Additionally, this method implements a more complex update cycle that recalculates cluster centroids after each individual reassignment.

These variations of K-Mean methods were chosen for a number of reasons. Firstly, K-Means++ offered insight into a more intelligent initialization method, while using the same clustering process as the original K-Means. This enabled us to consider how the performance differs based purely on differences in initialization methods. Secondly, Hartigan Wong K-Means used a different initialization method, as well as a slightly different clustering method, providing insight into how more advanced K-Means implementations perform. The variation between these two methods offered an opportunity to better understand the underlying factors that contribute to the performance

of a K-Means clustering, and additionally develop a sense for the application of different implementations.

To compare these methods, we designed a simple experiment that considered the mean performance for each method across a range of optimal and non-optimal k-values. For this study, we quantified the performance of a clustering using the F-β metric for $\beta = 1$. This allowed us to consider both the precision and recall of each method, offering a more holistic quantification of performance while remaining standardized and easy to compare. This metric was computed for each of the three methods across a range of k-values. Using the elbow plots presented in Figure 1, we concluded that the most optimal number of clusters in these data were three, with the Heart Wong algorithm perhaps favoring four or five clusters. Using this information, we set the range of k-values to be between one and six; a small enough set to accommodate our limited research efforts, but still large enough to provide insight into the performance envelope for each method.

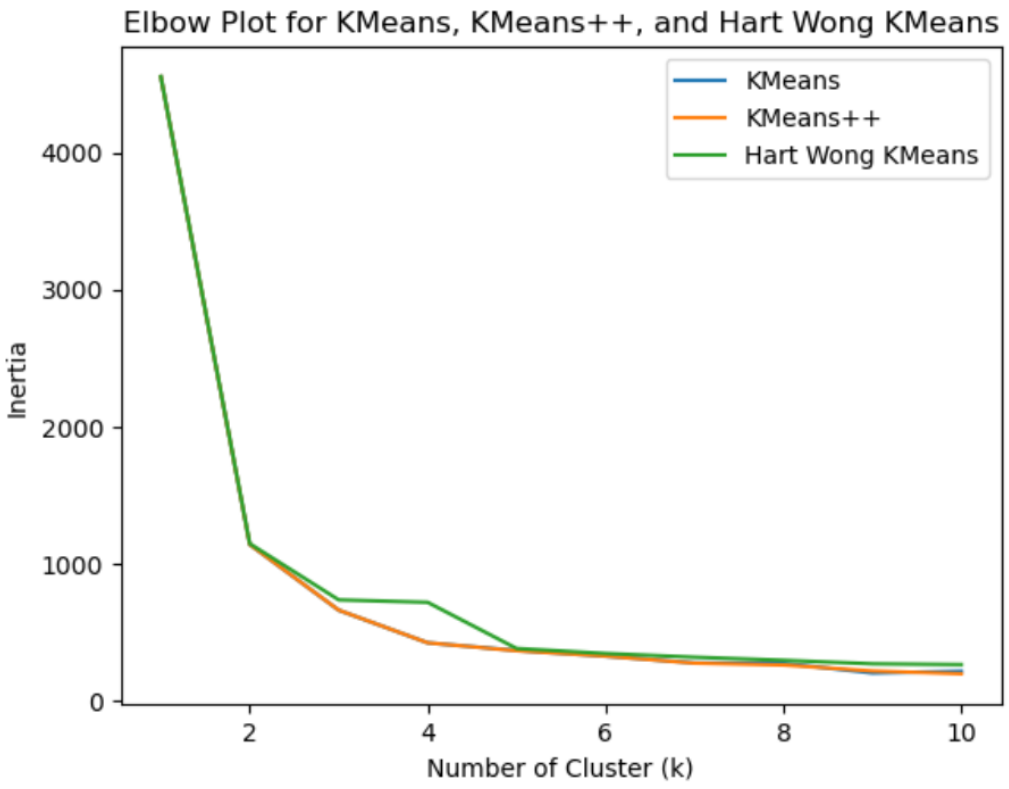

Figure 1. Elbow Plot for K-Means, K-Means++, and Hart Wont K-Means

With this range of k-values defined, we then fit clusters to these data using each of the three methods and recorded the resulting F-1 scores. This was repeated 100 times for each permutation of K-Means implementation and k-value, yielding the average performance of each method across a range of clustering conditions. Additionally, we generated scatter plots for each permutation, which allowed us to consider differences in the underlying clustering behavior of these methods. As explored in the following section, these results were used to create visualizations that supported our analysis efforts and allowed us to address our initial research question.

Results

The results from our experiment yielded very surprising results. In particular, the K-Means++ method proved to be much more resistant to less-optimal values of k. This is clearly demonstrated in Figure 2, where the mean F-1 score for the K-Means++ algorithm appears to remain relatively large across the range of k-values considered. Additionally, this method notably outperforms both the traditional K-Means and Hartigan Wong K-Means implementations, indicating that K-Means++ was the most optimal method for these data. This leads us to believe that K-Means++ could be more useful in situations where there is limited a-priori knowledge about the number of clusters in a dataset.

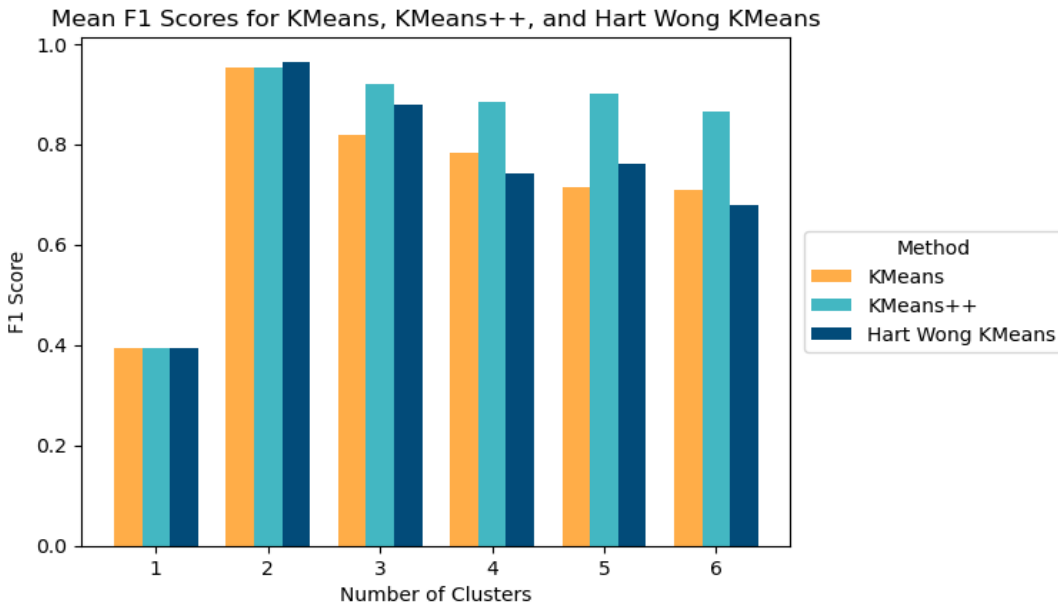

Figure 2. Bar chart of F1 scores.

Because the only difference between the traditional K-Means and K-Means++ algorithm was centroid initialization techniques, we believe that this resistance to non-optimal k-values is the direct result of a more intelligent initialization process. That said, it is surprising that the Hartigan Wong method, which also implements a more advanced initialization technique, did not perform similarly to the K-Means++. Instead, the Hartigan Wong method appears to provide slightly better F-1 scores for more optimal values of k (around k = 2 and k = 3), but quickly converges to the same performance as K-Means as k becomes less optimal. Once again, this is contrary to our initial expectations, where we predicted the strategic initialization and increased update rate of the Hartigan Wong method would yield better average performance than K-Means and K-Means++. We suspect that this deviation in expected behavior is the result of the initialization technique employed by Hartigan Wong, which still relies on randomly placing initial centroids. Future research might consider using different initialization methods for this algorithm to better understand the duality between cluster assignment frequency and centroid initialization.

In addition to differences in performance, we also consider the differences in cluster behavior for each of the implementations. Generally, K-Means and K-Means++ tended to produce very similar cluster assignments. This was expected due to both algorithms implementing the same underlying clustering approach. That said, K-Means++ tended to produce bigger clusters for larger values of k, as shown in Figure 4 (notice that K-Means produced two clusters around (5, 0) while K-Means++ produced only one). This behavior demonstrates that K-Means++ tends to be better at preventing multiple centroids from covering to the same local minima, confirming our previous observations.

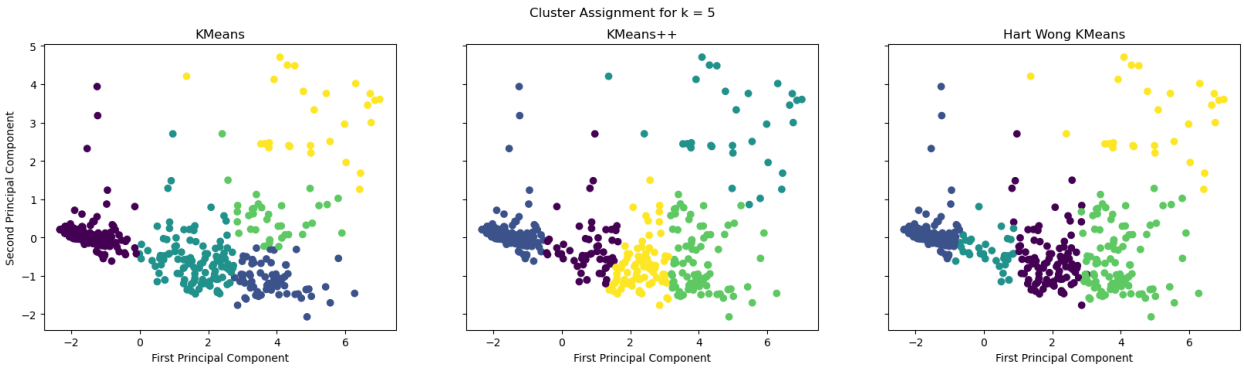

Figure 3. Scatter plots of K-Means clusters with k = 5.

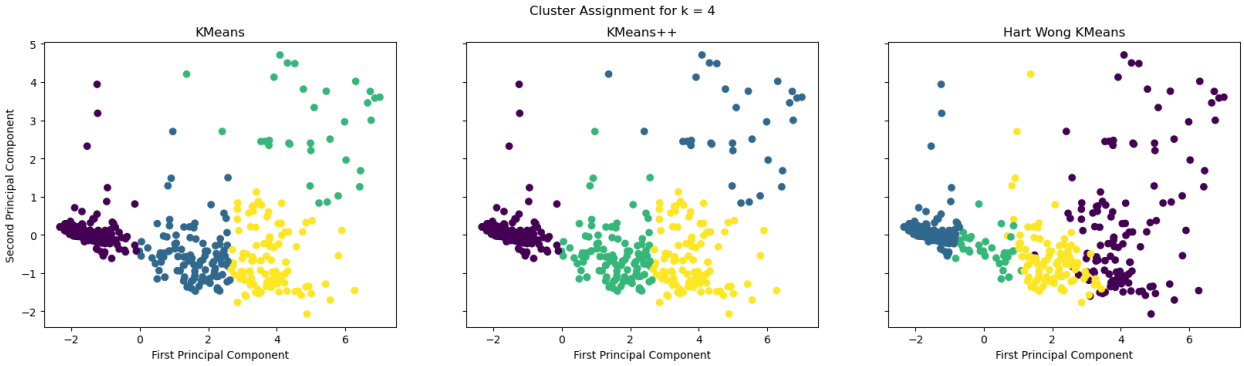

Figure 3. Scatter plots of K-Means clusters with k=4.

The Hartigan Wong method also demonstrates interesting clustering behavior, performing similarly to K-Means++ for larger k-values but producing much less cohesive clusters for lower k-values. As demonstrated in Figure 3, the K-Means and K-Means++ methods both identified two clusters for data whose First Principal Component value was greater than 2, separating two regions with very different densities. The Hartigan Wong method however, did not handle this difference in density. Instead, it combined both regions into a single cluster. Though we are uncertain of why this behavior manifests itself in the Hartigan Wong method, we suspect it is once again an artifact of the increased frequency with which cluster centroids are updated. Further research might more carefully consider differences in clustering behavior for these methods, especially in the context of more varied data.

Limitations and Study Considerations

Although the results of our study proved to be quite interesting, they should be carefully considered before producing any conclusions. Perhaps most notably, our experiment only considered a single dataset, representing a very specific context. This severely harms the generalizability of our results as certain features in the Breast Cancer dataset could have impacted the performance and behavior of the chosen K-Means implementations. Additionally, the performance of these methods might vary significantly depending on the structure of data, such as cluster shape, density, and proximity. Additional research should be conducted on a wide range of highly varied datasets to develop a more complete understanding of how these methods compare in different contexts.

Another major limitation of our study was that we only considered two alternative K-Means implementations. Once again, this constrains our results to only the K-Means++, Hartigan Wong K-Means, and similar implementations. The small variability between these two methods additionally limits our understanding of the different factors that contribute to clustering performance and behaviors. That said, given more time to perform this study we would consider a wider and more variable range of K-Means implementations to develop a more comprehensive understanding of the factors that yield improvement over the traditional algorithm. Additionally, we would consider these methods in the context of a many diverse datasets to better understand how data features affect clustering performance and behavior.

Conclusion

In conclusion, the observed results from our study indicate that there are alternative implementations of the K-Means algorithm that mitigate non-optimal convergence of centroids. In particular, the K-Means++ implementation appears to be much more robust to these situations, indicating that initialization techniques play a fundamental role in preventing centroids from getting caught in local minima. Additionally, the variable performance of the Hartigan Wong method indicates that specific implementations might be better suited for specific situations; for example, Hartigan Wong might be most useful when the parameters for the data are known while K-Means++ may be more applicable when little a-priori knowledge on the underlying data. That said, these conclusions remain incomplete as the many limitations of our study warrant further research into the impact of non-optimal centroid convergence of K-Means, and the possible solutions to this problem.

References

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

Probabilistic extension of precision, recall, and F1 score for more ... (n.d.). Retrieved May 6, 2023.