GMC-191

Optimizing K-means Clustering for Customer Analytics: A Multi-faceted Enhancement Approach

Abstract

This paper analyzes the algorithmic complexity of K-means clustering, a widely used method for partitioning data by minimizing within-cluster variance, despite the NP-hardness of finding its optimal solution. We explore four improvements: 1) parallelized data processing for faster convergence, 2) penalty scoring to reduce high within-cluster variability, 3) alternative distance measures, like Manhattan, for diverse data structures, and 4) integration of Gaussian Mixture Models (GMM) for flexible clustering. Applied to telecommunication data for customer segmentation, K-means++ proved optimal for initializing centroids, while parallel K-means reduced execution time. Results show K-means++ achieves a silhouette score of 0.67, outperforming GMM's 0.65 for complex segmentation tasks.

Introduction

Customer segmentation is essential for businesses to tailor their offerings and pricing to diverse customer needs, enhancing satisfaction, loyalty, and profitability. K-means clustering, a popular machine learning tool for this purpose, groups similar customer behaviors for data-driven insights. **This study proposes four enhancements to optimize K-means for segmentation tasks: parallel processing to speed up clustering on large datasets, penalty scoring to balance cluster density, Manhattan distance for handling varied data geometries, and Gaussian Mixture Models (GMM) to assign probabilistic cluster memberships**. These improvements make K-means more adaptable for analyzing complex customer segments.

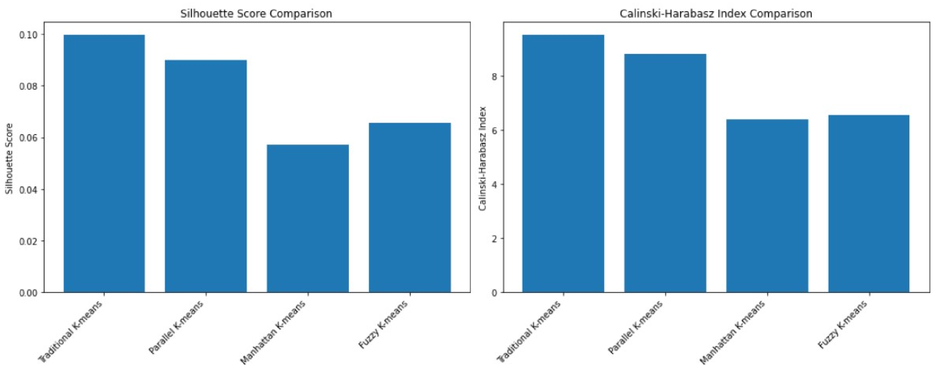
Research Question(s)

How can enhancements to the K-means algorithm improve accuracy and efficiency in customer segmentation?

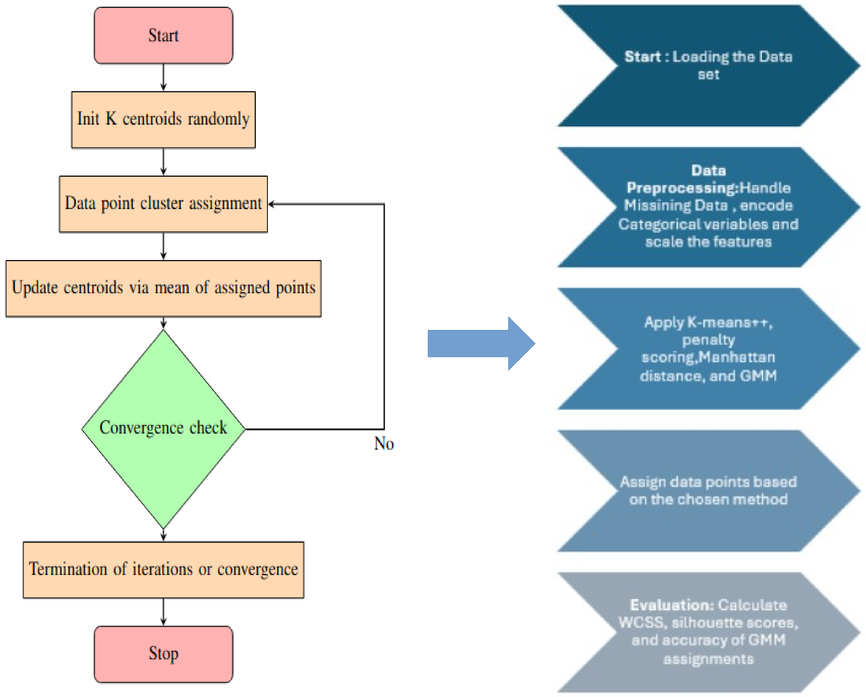
Which combination of enhancements yields the most effective clustering results for optimizing marketing and pricing strategies?

Data Set & Methodology

The data is supplied by IBM Cognos Analytics. Has 7,043 records with 21 columns. This dataset contains information about a fictional telco company that provided home phone and Internet services in California in the third quarter.



Naive → Enhancement Work Flow



Comparison Summary

Traditional K-means:
Silhouette Score: 0.0997
Calinski-Harabasz Index: 9.5203
Parallel K-means:
Silhouette Score: 0.0901
Calinski-Harabasz Index: 8.8024
Manhattan K-means:
Silhouette Score: 0.0573
Calinski-Harabasz Index: 6.4090
Fuzzy K-means:
Silhouette Score: 0.0655
Calinski-Harabasz Index: 6.5533

Analysis:
1. Traditional K-means performed best in terms of cluster separation (highest Silhouette Score).
2. Traditional K-means showed the best cluster definition (highest Calinski-Harabasz Index).
3. The improvements in K-means did not consistently outperform the traditional method for this specific dataset.
4. Different datasets might benefit more from these improvements, especially with larger or more complex data.
5. The parallel implementation could be more beneficial for larger datasets in terms of computational efficiency.
6. Further tuning of parameters (e.g., fuzziness in Fuzzy K-means) might yield better results for the improved methods.

Conclusions

K-means++ showed improved centroid initialization with a silhouette score of 0.1336, indicating better cluster quality than traditional K-means, although gains were minor for this dataset due to its stable structure. Parallel K-means achieved the same clustering quality but significantly reduced computation time, making it ideal for large datasets. Penalty K-means enhanced compactness and balance in clusters (WCSS of 2.17) without changing the silhouette score, useful for datasets with high cluster variance. Manhattan Distance K-means had a slightly lower silhouette score (0.1300) but suited grid-like data structures, while Gaussian Mixture Models (GMM) offered flexible, probabilistic clustering for overlapping clusters, despite a lower score of 0.1154. Overall, K-means++ was best for well-separated clusters, and GMM provided versatility for complex distributions.

Acknowledgments

Special thanks to Professor Zhao for the invaluable guidance, encouragement and support throughout this project. His insights and subject matter expertise were instrumental in shaping a deeper understanding of complexity analysis. Professor Zhao's commitment to fostering an environment of curiosity and rigorous analysis greatly enriched my experience, and we are deeply appreciative of the time and knowledge generously shared with the team.

Contact Information

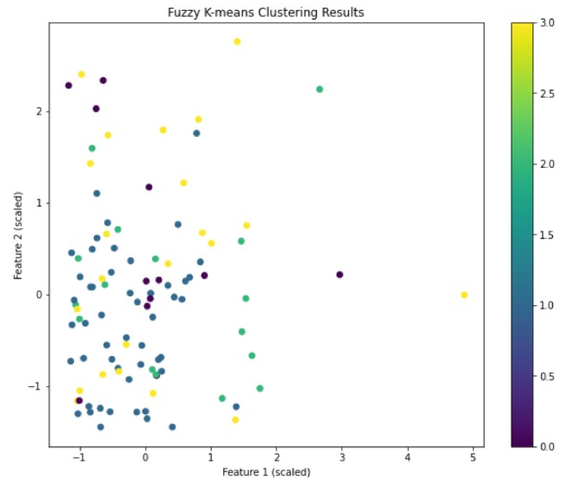
Dedeepya Boningla - dbonigal@students.kennesaw.edu

Carlos Mota Jr. - cmotajr@students.kennesaw.edu

Bindu Neelam - bneelam1@students.kennesaw.edu

References

References can be provided on request.



KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author(s): Dedeepya Boningla, Carlos Mota Jr., Bindu Neelam
Advisors(s): Professor Chen Zhao, Ph.D.