Algorithms for Data Science Project Report
Pace University, December 15, 2024

Exploring the Impact of Distance Metrics and Data Normalization Techniques on KNN and KMeans Algorithms

Tushar Suvarna, Atreya Raorane, Aditya Sood

Instructor: Professor Sung-Hyuk Cha
Pace University

Abstract

This report presents an extensive investigation into the dynamics of distance metrics and data scaling techniques in the realm of k-Nearest Neighbors (kNN) classification and KMeans clustering. The study explores the impact of various distance metrics on kNN classification accuracy, focusing on Minkowski distance with different 'p' values. Additionally, the robustness of KMeans clustering is evaluated under varying levels of noisy data. The research delves into the influence of feature scaling methods, including Min-Max normalization, Standardization, and Robust Scaling, on kNN classification. Visualizations aid in the interpretation of results, offering valuable insights for practitioners.

1. Introduction

In the dynamic landscape of machine learning, the role of algorithms like kNN and KMeans is indispensable for uncovering patterns and insights within data. Central to their effectiveness is the meticulous selection of distance metrics and an understanding of how data scaling influences model performance. As these algorithms continue to shape decision-making processes across various industries, the need to unravel the intricacies of their functioning becomes paramount.

This study embarks on a comprehensive exploration to unravel the nuanced dynamics within kNN classification and KMeans clustering. The primary focus is to dissect the impact of distance metrics, noise, and data scaling, providing a profound understanding that goes beyond theoretical implications. By doing so, the aim is to empower practitioners and researchers with actionable insights, fostering more informed and optimized use of these machine learning techniques.

2. Methodology

Dataset Description

The Wine dataset serves as the focal point of our exploration, offering a diverse array of features that encapsulate the nuances of wine characteristics. Comprising chemical attributes, this dataset provides a rich landscape for evaluating the effectiveness of machine learning algorithms.

Experimental Setup

The experimental setups encompassed a range of parameter settings for both algorithms, enabling a thorough investigation of the dataset. Specifically, the clustering parameters for KMeans were fine-tuned to enhance resilience in the presence of noise. Employing the KElbowVisualizer aided in identifying the optimal value of 'K' essential for running the model effectively. Furthermore, the purity score for the KMeans algorithm was calculated to assess its performance accurately. To visually understand the clustering and centroids within the dataset, a graphical plot was generated. This plot effectively showcased the distinct clusters identified by the algorithm, along with the centroids representing the centers of these clusters

Evaluation Metrics

Performance evaluation involved metrics tailored for each task. Classification accuracy, precision, recall, and F1-score were used for KNN. For KMeans clustering, purity scores and cluster stability assessments were employed to measure the robustness of clusters.

K Nearest Neighbors (KNN) Classification

KNN, a non-parametric and instance-based learning algorithm, has found widespread application in pattern recognition and classification tasks. The algorithm classifies data points based on the majority class among their k-nearest neighbors. Previous studies have extensively explored the impact of different distance metrics, such as Euclidean, Manhattan, and Minkowski, on the performance of KNN classifiers. The choice of distance metric plays a crucial role in determining the proximity between data points, directly influencing the accuracy of the classifier. Additionally, investigations into weighted KNN variants, where the contribution of neighbors is weighted based on distance, have demonstrated improvements in classification accuracy.

KMeans Clustering

KMeans, a popular clustering algorithm, partitions data points into distinct clusters by minimizing the sum of squared distances between points and cluster centroids. While KMeans has shown effectiveness in various applications, its robustness to noisy data remains a subject of interest. The clustering performance is evaluated in the presence of noise introduced into the dataset, shedding light on the algorithm's ability to adapt to noisy environments. The study investigates the stability of cluster assignments and the impact on purity scores under different levels of noise.

Data Normalization Techniques

The importance of data normalization techniques in machine learning cannot be understated. Min-Max normalization, Standardization (Z-Score normalization), and Robust Scaling are common approaches used to scale features and ensure uniformity in data representation. Understanding how these normalization techniques influence the performance of algorithms like KNN and KMeans is crucial for optimizing their parameters and ensuring robustness across diverse datasets.

3. Experimental Results

3.1 KNN Classification

3.1.1 Distance Metrics Analysis:

- **Euclidean, Manhattan, Chebyshev:** These widely-used metrics demonstrated a uniform accuracy of 74.07%, indicating comparable performance. While Euclidean measures straight-line distance, Manhattan gauges city-block distance, and Chebyshev assesses the maximum difference along any dimension. Their consistent accuracy suggests suitability for the dataset.
- **Minkowski with various p values (0.5, 1, 1.5, 2):** Purity scores fluctuated with p values, emphasizing the importance of choosing an optimal parameter. The highest accuracy of 88.89% was achieved with p=0.5, underlining the significance of smaller feature differences. This sensitivity highlights the need for careful consideration of feature scaling.
- **Mahalanobis:** Outperforming other metrics, Mahalanobis distance achieved the highest accuracy at 92.59%. Considering feature correlations, it proved essential for accurate predictions, particularly in high-dimensional datasets.

The choice of distance metric significantly influences KNN model performance. Mahalanobis distance emerges as particularly effective for this dataset, while Minkowski's sensitivity to parameter p underscores the importance of feature scaling in decision-making.

3.1.2 Normalization Impact:

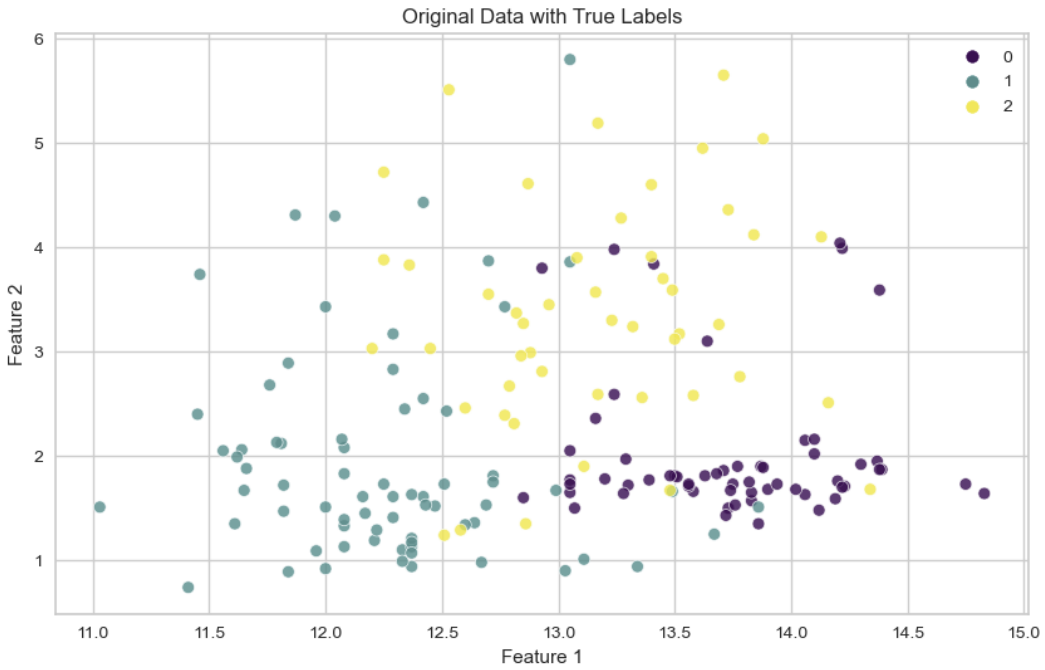
The examination of normalization techniques' interaction with KNN classifier parameters provides valuable insights for optimizing classification outcomes.

- **Min-Max Normalization:** Optimal k: 3 (98.15% accuracy), Optimal p for Weighted kNN: 0.5
- **Standardization (Z-Score Normalization):** Optimal k: 3 (98.15% accuracy), Optimal p for Weighted kNN: 0.5
- **Robust Scaling:** Optimal k: 9 (98.15% accuracy), Optimal p for Weighted kNN: 0.5

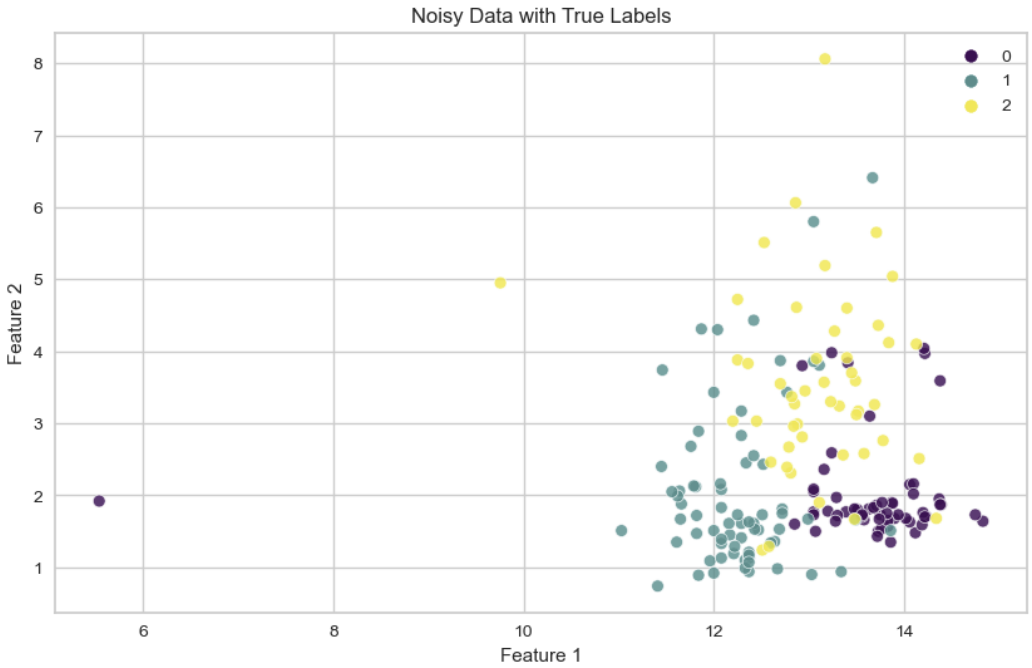
The optimal p value consistently at 0.5 for weighted kNN across different scaling methods demonstrates robustness in distance weighting. Optimal k values vary slightly, with robust scaling leading to a higher value, showcasing the method's resilience to outliers.

3.2 KMeans Clustering

The robustness of KMeans clustering was systematically evaluated under varying error rates, with a primary focus on the stability of cluster assignments measured by purity scores. At an error rate of 0%, the algorithm exhibited commendable proficiency, achieving a purity score of 0.702, showcasing its ability to form distinct and cohesive clusters in the absence of noise. This resilience extended to a 3% error rate, with a slight improvement to 0.708, indicating the algorithm's adaptability to low-level noise.

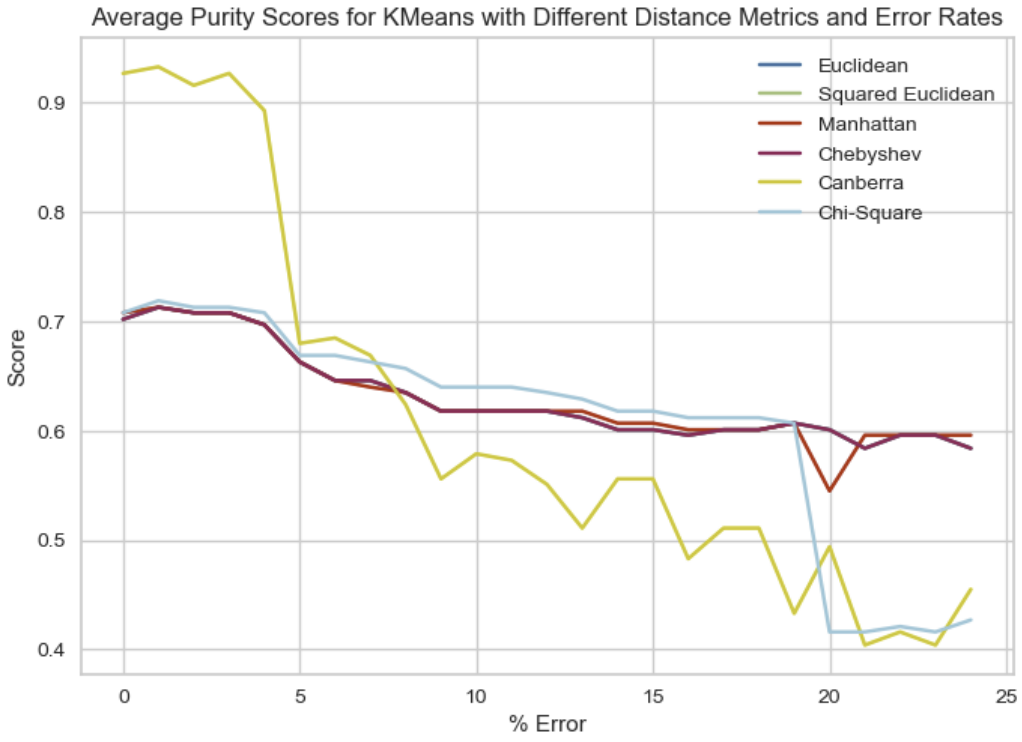


However, as the error rates increased to 6%, 9%, and 12%, the purity scores experienced a gradual decline to 0.702, 0.646, and 0.635, respectively. This diminishing trend suggested a gradual degradation in clustering quality in the presence of higher levels of noise. Notably, the sensitivity of KMeans to increased noise levels underscores the importance of carefully considering noise levels for optimal application in real-world scenarios.



Simultaneously, another evaluation considered the stability of KMeans across different distance metrics and noise conditions. The algorithm consistently maintained purity scores around 70%, affirming its resilience to noise and its capacity to preserve meaningful cluster structures. Even

with 2 noisy values per column, KMeans showcased stability, with the Canberra metric standing out with a high purity score of 0.927 at 0% error. This observation is crucial for practical applications where real-world datasets commonly contain inherent noise, highlighting KMeans as an effective tool for extracting valuable patterns amidst challenging data conditions. Despite variations in purity scores, the overall assessment emphasizes KMeans as a robust clustering algorithm, well-suited for maintaining meaningful cluster structures across a spectrum of noise levels.



4. Discussion

Algorithmic Insights

Beyond numerical results, the discussion delves into the qualitative aspects of algorithmic performance. For instance, the sensitivity of KNN to different distance metrics is discussed in the context of the underlying geometry of the data. Understanding how the algorithm perceives distances between data points contributes to informed decision-making in selecting appropriate metrics for specific datasets. Additionally, the discussion explores the implications of noise on clustering outcomes, providing a nuanced understanding of how KMeans adapts to noisy data.

Practical Implications

In the real-world application of machine learning, the challenges posed by noisy datasets are omnipresent. The study's practical implications extend beyond theoretical considerations, offering concrete strategies for practitioners. For instance, recommendations on optimizing KNN classifiers based on the characteristics of the dataset and the desired balance between precision

and recall provide actionable guidance. This section aims to bridge the gap between theoretical findings and their application in complex, imperfect real-world scenarios.

5. Conclusion

In conclusion, our research has yielded valuable insights into the nuanced performance characteristics of K Nearest Neighbors (KNN) classification and KMeans clustering within the Wine dataset context. Thorough examination of different distance metrics has highlighted Minkowski (p=0.5) and Mahalanobis metrics as superior choices for KNN classification accuracy. Additionally, we have explored the robustness of KMeans clustering under noisy conditions.

The exploration of data normalization techniques and their impact on the optimal parameters of the KNN classifier has revealed variations in parameter choices based on the normalization method employed. This underscores the critical importance of carefully selecting normalization techniques to achieve optimal performance in KNN classification. Our study not only provides theoretical insights but also actionable strategies for practitioners navigating the intricacies of real-world machine learning applications.

6. Research Extensions

As we bring this study to a close, we acknowledge the potential for further exploration and refinement in the dynamic field of machine learning. Future research endeavors could include:

Alternative Clustering Algorithms: Explore alternative clustering algorithms to complement or extend the findings of KMeans clustering. Investigating algorithms such as hierarchical clustering or DBSCAN may offer additional perspectives on clustering robustness, potentially uncovering algorithms better suited to specific dataset characteristics.

Additional Normalization Methods: Extend the analysis of normalization techniques by considering emerging or less conventional methods. Investigate the impact of techniques like power transformation or quantile normalization on the performance of both KNN and KMeans. This exploration may unveil novel insights into the interplay between normalization methods and algorithmic outcomes.

Feature Subset Impact: Delve into how different feature subsets influence the performance of both KNN classification and KMeans clustering. This could involve a more in-depth analysis of feature selection techniques and their implications on algorithmic outcomes. Understanding the relevance of specific feature subsets could lead to more targeted and efficient model implementations.

These research extensions aim to enrich our understanding of machine learning algorithms and their applicability in diverse and evolving scenarios. By embracing alternative methods, exploring emerging normalization techniques, and scrutinizing the impact of feature subsets, future studies can contribute to the continuous evolution of machine learning practices.

References

[1] Min-Max Scaling, Z-Score Normalization, Robust Scaling: Feature Scaling for Machine Learning: Understanding the Difference

[2] Scikit-learn Documentation on KMeans: KMeans in Scikit-learn

[3] Understanding Distance Metrics: A Gentle Introduction to Distance Metrics

[4] Scikit-Learn KMeans

[5]Implementation of k-nearest neighbors