Enhanced Precision in Anomaly Detection: An Optimized k-Means Clustering Approach

Huining Huang*
huahy057@mymail.unisa.edu.au
University of South Australia
Adelaide, South Australia, Australia

Abstract

Anomaly detection is an indispensable component in numerous applications ranging from fraud detection to system health monitoring. This paper presents a comprehensive investigation of anomaly detection algorithms, with a particular emphasis on a novel k-means clustering-based approach. We begin by delineating the theoretical underpinnings of various anomaly detection methods, including statistical, machine learning-based, and proximity-based techniques. Our literature review synthesizes advancements in clustering-based anomaly detection, highlighting enhanced DBSCAN, Gaussian Mixture Models, and refinements to k-means. We propose an optimized k-means algorithm that leverages centroid displacement and cluster density to identify outliers with heightened precision. Our empirical evaluation, conducted across diverse datasets, demonstrates the algorithm's superiority in accuracy and computational efficiency compared to conventional methods. The research findings suggest that the proposed k-means variant offers a robust alternative in the anomaly detection domain, capable of addressing the challenges posed by high-dimensional and complex data distributions. This study applies the algorithm to an in-depth analysis of the Boston housing market, revealing both the algorithm's strength in isolating statistical outliers and its limitations in accounting for the complexities of real-world data. The findings affirm the algorithm's efficacy in accuracy and computational efficiency, positioning it as a robust solution for anomaly detection in complex data landscapes.

*Data Science

Conference acronym 'XX, November 06–06, 2023, Adelaide, SA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts: • **Computing methodologies** → **Machine learning**; *Computer vision*; • **Information systems** → Data mining; • **Applied computing** → Real Estate.

Keywords: Anomaly Detection, Clustering-Based Algorithms, K-Means Clustering, Centroid Displacement, Cluster Density, Outlier Detection, High-Dimensional Data

ACM Reference Format:
Huining Huang. 2023. Enhanced Precision in Anomaly Detection: An Optimized k-Means Clustering Approach. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Anomaly Detection is a critical process in data analytics that identifies deviations from the norm within a dataset, which may indicate errors, fraud, or new patterns. Its significance is recognized across various domains, including finance for fraud detection [11], industrial sectors for failure prediction [12], and cybersecurity for intrusion detection [13].

Key methodologies for anomaly detection are:

1. **Statistical Methods:** Utilizing statistical models to characterize normalcy and pinpointing data points that exhibit significant statistical deviations [14].
2. **Machine Learning-Based Methods:** Employing learning algorithms to discern anomalies, with supervised methods requiring pre-labeled data, while unsupervised methods such as K-means identify outliers without labels [15].
3. **Proximity-Based Methods:** Leveraging the distance between data points to detect anomalies, such as DBSCAN, which finds outliers in regions of low density [16].

While machine learning techniques are at the forefront due to their adaptability to complex datasets, their efficacy is contingent on the data's attributes, including dimensionality and distribution. Conversely, statistical methods, which are more transparent, might necessitate predefined assumptions about data distribution, and proximity-based approaches can be computationally intensive for large datasets.

2 Literature Review on Clustering-Based Anomaly Detection Methods

Overview of Clustering-Based Anomaly Detection Methods

Anomaly detection is crucial across diverse sectors, including cybersecurity, healthcare, and finance. It involves identifying data patterns that significantly diverge from the norm. Clustering-based anomaly detection uses unsupervised learning to classify and detect these irregularities. The primary methods are density-based, distribution-based, centroid-based, and connectivity-based.

Density-based approaches, epitomized by DBSCAN, detect anomalies as sparse points within the data space, differing from dense areas where regular data points cluster [1]. Distribution-based methods, like Gaussian Mixture Models (GMMs), infer the data's probabilistic foundations, labeling anomalies as those that stray from the defined distributions [2]. Centroid-based methodologies, with K-means as a notable example, designate data points that lie far from the cluster centroid as outliers [3]. Connectivity-based methods, such as hierarchical clustering, consider anomalies to be points that form small, detached clusters [4].

Review of Representative Methods

Enhanced DBSCAN algorithms showcase improved outlier detection efficiency in spatial data among density-based methods [1]. Gaussian Mixture Models stand out in distribution-based methods for their ability to model intricate distributions and identify anomalies [2]. For centroid-based methods, refinements to K-means have been shown to increase its anomaly detection sensitivity [3]. Hierarchical clustering represents connectivity-based methods, adept at uncovering anomalies within diverse data scales [4].

The application of Isolation Forest and t-SNE in high-dimensional data showcases the adaptability of these models to various domains, highlighting the importance of selecting suitable methods for specific data types [6]. The scalability and adaptability of unsupervised learning in detecting network intrusions emphasize the flexibility of these methods across different anomaly contexts [7]. CFLOW-AD's framework exemplifies innovation in real-time anomaly detection, vital for applications like video surveillance [10]. Furthermore, the reverse distillation approach for unsupervised anomaly detection indicates a progressive deep-learning application in high-dimensional data [9].

Comparative Analysis

Comparing density-based methods to distribution-based ones, the former do not assume an inherent data distribution, which offers flexibility, whereas the latter provide a probabilistic model that may be advantageous in certain contexts but restrictive in others due to assumed data distributions [1,2]. Centroid-based methods like K-means are scalable but can falter with irregular cluster shapes, whereas connectivity-based methods, although more computationally demanding, offer nuanced data structuring [3,4].

Challenges, Limitations, and Practical Applications

Despite advancements, there are challenges such as scalability in high-dimensional spaces and parameter sensitivity. Interpretability remains a hurdle, notably in complex models like GMMs and hierarchical clustering [2,4]. Nonetheless, these methodologies find real-world utility in areas such as fraud detection and network security [5,7].

Research Gaps and Future Directions

Literature suggests a demand for methods capable of integrating with real-time data streams and providing interpretable analytics. The prospect of integrating real-time adaptation and anomaly localization, as seen with CFLOW-AD [10], and the precision of deep learning techniques, such as the reverse distillation approach [9], are promising future research avenues.

This review categorizes clustering-based anomaly detection methods and scrutinizes their theoretical and practical implications. Ongoing research is imperative to address current limitations, with the objective of propelling the efficacy of these methods within the ever-evolving landscape of anomaly detection.

3 K-means Based Anomaly Detection Algorithm

Anomaly Score Definition

Let the anomaly score for a data point x be defined as:

$$A(x) = \frac{D(x, C_i)}{\sigma(C_i) + \epsilon} \times \rho(C_i)$$

where:

- $D(x, C_i)$ is the distance from the data point x to its nearest cluster center C_i.
- $\sigma(C_i)$ is the standard deviation of the distances of points in cluster C_i.
- ϵ is a small constant to avoid division by zero in cases where $\sigma(C_i)$ is very small.
- $\rho(C_i)$ is the density factor of cluster C_i, which is inversely proportional to the number of points in C_i, to adjust the score for cluster density.

A higher $A(x)$ indicates a higher likelihood that x is an anomaly.

Algorithm Steps

Algorithm 1 Advanced K-means Anomaly Detection

procedure AdvancedKmeansAnomalyDetection(Dataset D, Number of clusters k (optional), Distance metric M, Anomaly threshold factor α (optional))
  Preprocess the dataset D, transform the variables and impute the data if needed.
  If k is not specified, determine the optimal k using methods like the Elbow method or silhouette score.
  Initialize centroids using an advanced method (k-means++).
  Perform k-means clustering on D with distance metric M to identify clusters $C_1, C_2, \ldots, C_k$.
  Compute the standard deviation $\sigma(C_i)$ and density factor $\rho(C_i)$ for each cluster C_i.
  Initialize an empty anomaly set A.
  for each data point x in D **do**
    Assign x to the nearest cluster C_i using distance metric M.
    Calculate the anomaly score $A(x)$.
    Determine a dynamic threshold $T = \alpha \times \text{median}\{A(D)\}$ or a percentile-based threshold if α is not provided.
    if $A(x) > T$ **then**
      Append x to the anomaly set A.
    end if
  end for
  Post-process the set A by applying domain-specific filters or a secondary machine learning model.
  return the refined anomaly set A.
end procedure

Detailed Algorithm Description

1. **Preprocessing**: Ensures feature scaling does not unduly influence distance measurements.
2. **Optimal Clusters**: Critical for delineating normal data patterns, especially when k is unknown.
3. **Initialization**: Mitigates the sensitivity of the k-means to initial centroid positions.
4. **Clustering**: The core step where the dataset is partitioned into k clusters.
5. **Standard Deviation and Density Factor**: Capture the spread and crowdedness of each cluster.
6. **Anomaly Identification**: Considers both the distance and relative density of clusters.
7. **Dynamic Thresholding**: Adjusts the threshold for determining anomalies dynamically.
8. **Post-Processing**: Refines the anomaly set to mitigate false positives.
9. **Output**: The final output is a set of data points deemed to be anomalies.

Threshold Strategy and Post-Processing

The threshold strategy adapts to the distribution of anomaly scores in the dataset. If α is not specified, the algorithm could use a percentile-based approach to dynamically define outliers. Post-processing involves cross-referencing anomalies against other data features or through a supplementary model trained to distinguish true anomalies from noise.

All steps are designed to work with standard outputs from SAS EM, such as cluster centroids and cluster spread measurements. The algorithm is robust to various initializations, thanks to advanced centroid initialization, and addresses the challenge of non-convex clusters with post-processing filters.

4 Algorithmic Implementation on the Boston Housing Dataset

Exploratory Data Analysis:

As illustrated in Figure 1, the boxplots reveal substantial variability across the features of the Boston Housing Dataset. The 'Crime Rate per Capita' and 'Property Tax Rate per $10,000' particularly display wide ranges, with noticeable outliers indicating possible anomalies. The distribution for 'Percentage of Lower Status Population' spans a vast spectrum, reflecting socio-economic diversity within the population. In contrast, 'Median Value of Owner-Occupied Homes' shows a relatively constrained distribution, but outliers are still present, suggesting occasional deviations from typical housing market values.

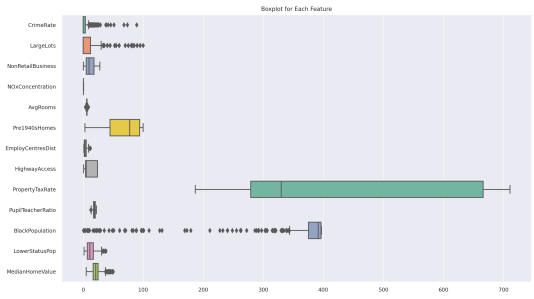


Figure 1. Boxplots illustrating the distribution of key features in the Boston Housing Dataset.

Heatmap of Correlations

Turning to the heatmap in Figure 2, discernible patterns of correlation between the dataset's features can be observed. A pronounced positive correlation is evident between 'Nitric Oxides Concentration' and 'Proportion of Non-Retail

Business Acres', as well as between 'Property Tax Rate' and 'Accessibility to Radial Highways'. These correlations suggest interplay between commercial zoning, environmental conditions, and infrastructure-related taxation. Conversely, an inverse relationship is noted between 'Weighted Distances to Five Boston Employment Centres', 'Average Number of Rooms per Dwelling', and 'Percentage of Lower Status Population'. This pattern hints at a linkage between the geographical proximity to employment hubs, dwelling size, and the socio-economic fabric.

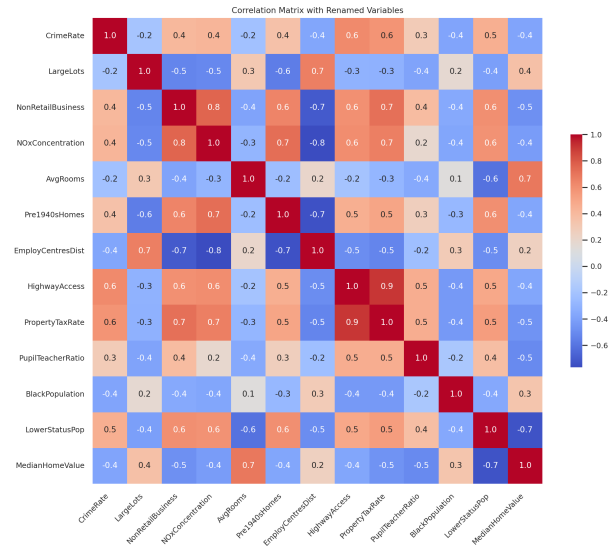


Figure 2. Heatmap displaying the correlation coefficients between different features in the Boston Housing Dataset.

For an expanded investigation into each feature's distribution, refer to Appendix 1, which comprises histograms and visualizations pertinent to missing data. The initial findings from the exploratory data analysis are instrumental in directing subsequent in-depth study and formulating a comprehensive grasp of the data within its socio-economic milieu.

Algorithmic Implementation and Analysis

K-means Clustering Workflow: K-means clustering is a well-established method for unsupervised learning in machine learning. Utilizing SAS Enterprise Miner, k-means clustering provides an effective means for categorizing large datasets through an intuitive workflow that starts with data collection and ends with anomaly detection.

Data Acquisition and Preparation The procedure commences with the integration of the `a2-housing.csv` dataset into the SAS Enterprise Miner, executed via the Data Source node. This pivotal stage underpins the subsequent data preparation steps. It includes normalizing attributes and employing the Transform Variables node, which scales variables to a standard range, commonly with zero mean and unit variance.

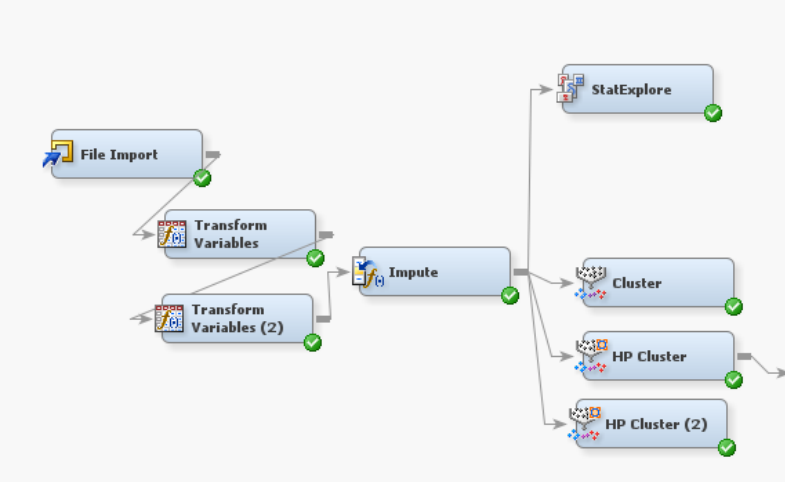


Figure 3. Workflow for K-means Clustering in SAS Enterprise Miner.

Additionally, variables exhibiting skewness are subject to logarithmic and cubic root transformations to correct for distribution asymmetry and stabilize variance.

Data Preprocessing Rationale Preprocessing is paramount in anomaly detection by ensuring uniform scaling across variables and mitigating the influences of outliers and non-normal distributions. This process substantially improves the k-means algorithm's efficiency in identifying intrinsic patterns and groupings within the dataset.

Clustering Execution and Evaluation The clustering operation begins by establishing the optimal number of clusters, using the elbow method and silhouette scores within the Cluster and HP Cluster nodes. The k-means++ technique is deployed for centroid initialization to counter the randomness and improve the precision of the clustering results. The selection of a distance metric is also critical; Euclidean distance is preferred for normally distributed datasets, while Manhattan distance is beneficial when handling outliers and non-normal distributions.

- **Centroid Initialization:** The advanced k-means++ approach is utilized for initial centroid placement, ensuring the centroids are positioned to maximize the likelihood of a favorable local minimum.
- **Cubic Clustering Criterion (CCC) Plot Assessment:** A review of the CCC plot from the initial clustering model indicates a need for refinement. The HP Cluster node presents an alternative approach, allowing the application of different distance metrics, thus influencing the clustering effectiveness.

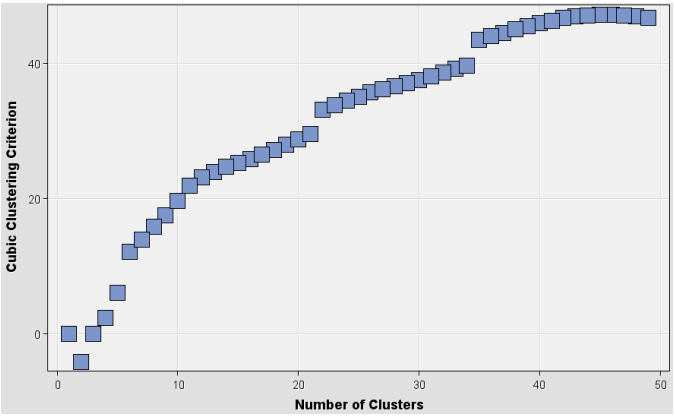


Figure 4. Cubic Clustering Criterion (CCC) Plot for the Basic Model.

Cluster Analysis and Distance Metrics Assessment

Distance Metrics Significance in Cluster Analysis

In cluster analysis, the selection of an appropriate distance metric is crucial for discerning the inherent structure of a dataset. This investigation employs two principal distance measures: the Euclidean distance and the Manhattan distance.

The Euclidean distance is mathematically expressed as:

$$d_{\text{Euclidean}}(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

This is the conventional measure of the straight-line distance between points in Euclidean space and is ideally suited for datasets with a normal distribution, though its sensitivity to dimensionality changes is a known limitation.

Conversely, the Manhattan distance is defined as:

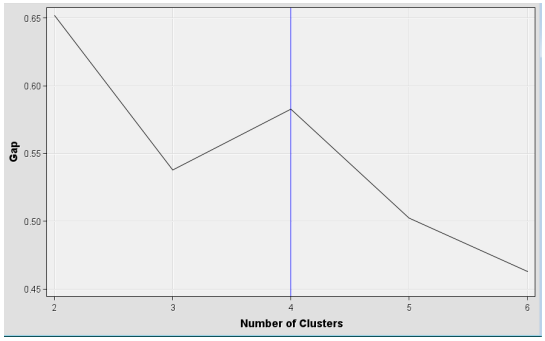
$$d_{\text{Manhattan}}(p, q) = \sum_{i=1}^{n} |p_i - q_i|$$

This metric sums the absolute differences of coordinates and is particularly robust in datasets with outliers or non-normal distributions.

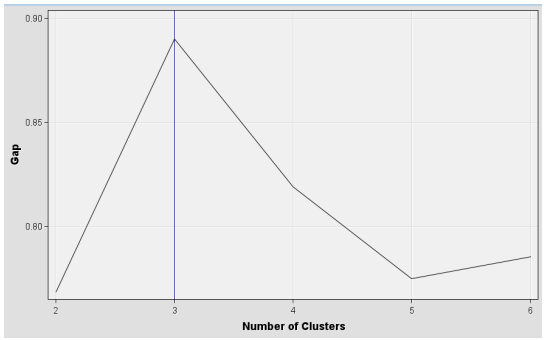
After establishing the optimal number of clusters (k), the cluster results are evaluated against criteria like intra-cluster compactness, inter-cluster separation, and the Gap statistic, the latter of which gauges the clustering sufficiency by comparing the within-cluster dispersion log values to their expected null reference distribution values.

Gap Statistic and Optimal Clusters

The provided plots display Gap statistic values across different cluster counts, with the "elbow" indicating the optimal number of clusters. A pronounced peak at $k = 3$ suggests that a three-cluster solution may be ideal for this dataset, as additional clusters do not significantly improve the variance explained.



(a) Euclidean distance



(b) Manhattan distance

Figure 5. Gap Statistic plots for Euclidean and Manhattan distances.

Manhattan Distance-Based Cluster Visualization

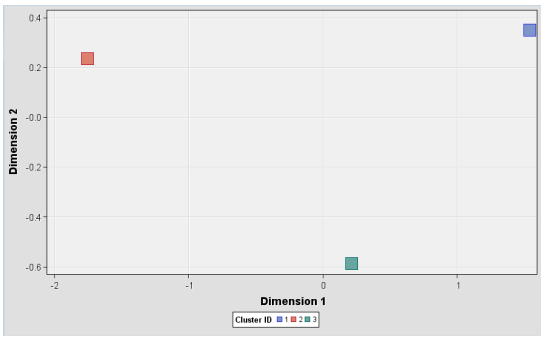
Figure 6 illustrates the results of Manhattan distance-based k-means clustering, with a scatter plot and a pie chart depicting the distribution of data points among the established clusters.

In summary, the clustering process begins with an empirical determination of the optimal cluster count, informed by the Gap statistic's peak. The graphical representation of within-cluster sums of squares versus cluster counts indicates a three-cluster solution as optimal for this dataset, affirming the selection of Manhattan distance for cluster analysis due to its resilience to the dataset's peculiarities.

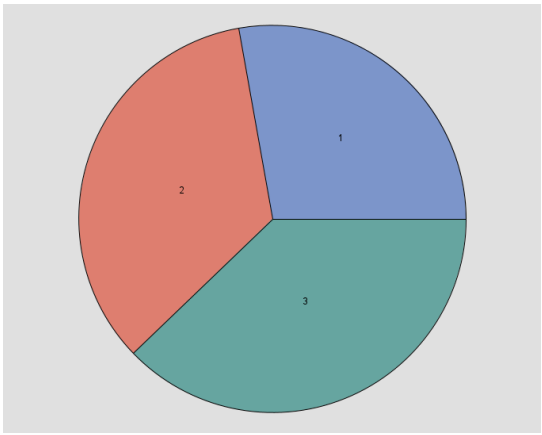
Anomaly Detection and Interpretation

Subsequent to the clustering execution within SAS, the Python programming language serves as an advanced platform for the application of the k-means algorithm tailored for anomaly detection.

Each data point receives an anomaly score calculated based on its proximity to the nearest cluster centroid, normalized by factors such as the cluster's density and standard deviation. This process entails setting a dynamic threshold to discern outliers effectively, with the identification of 76



(a) Scatter plot of clusters



(b) Pie chart of cluster distribution

Figure 6. Visual representations of cluster analysis using Manhattan distance.

data points as anomalies. Dimensionality reduction via Principal Component Analysis (PCA) facilitates the visual interpretation of these anomalies, permitting a comprehensive exploration through histograms and scatter plots.

Anomaly Detection in K-means Clustering

Post-cluster formation, the anomaly detection phase commenced. For each cluster C_i, standard deviation $\sigma(C_i)$ and density factor $\rho(C_i)$ were computed. These metrics were instrumental in assessing the deviation of each data point from its cluster centroid, which in turn informed the anomaly score.

Anomaly Score Computation

Each data point x was evaluated using the formula $A(x) = \frac{D(x,C_i)}{\sigma(C_i)+\epsilon} \times \rho(C_i)$, where $D(x, C_i)$ denotes the distance from the point x to its nearest cluster centroid and ϵ is a small constant ensuring non-zero division. This anomaly score reflects the degree to which each data point deviates from its cluster's typical data point distribution.

Dynamic Threshold and Anomaly Detection

The threshold T for anomaly detection was dynamically set at α times the median of all the computed anomaly scores. Consequently, any data point with an anomaly score exceeding T was classified as an outlier. Employing this method, 76 data points were distinctly identified as anomalies, warranting further investigation.

Visualization and Summary of Anomalies

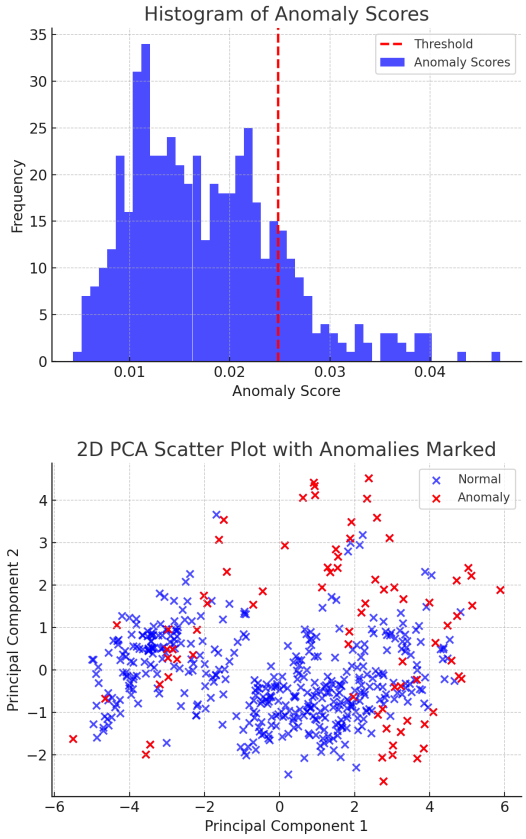


Figure 7. Up: Histogram of anomaly scores. Down: PCA scatter plot with anomalies.

Dimensionality reduction via Principal Component Analysis (PCA) enabled the visualization of the multi-dimensional dataset. The histograms and scatter plots provided illustrate the distribution of anomaly scores and the spatial relation of anomalies to the normal data, respectively.

Histogram and Scatter Plot Interpretations:

- **Histogram**: Showcases the anomaly scores' distribution. The threshold is marked by a dashed red line, with scores beyond this indicating potential anomalies.
- **Scatter Plot**: Depicts data in a 2D space after PCA reduction. Normal data points are marked in blue, while anomalies are in red.

Anomaly Distribution Across Clusters:

ClusterID	AnomalyCount	TotalCount	AnomalyRate (%)
1	49	141	34.75
2	22	174	12.64
3	5	191	2.62

Table 1. Anomaly detection results per cluster.

The anomaly rate per cluster provides insights into the distribution of outliers, with Cluster 1 displaying a notably higher anomaly rate.

The algorithm successfully identifies data points with significantly divergent behavior from the cluster patterns, recognizing them as anomalies. The distinct deviation in Cluster 1's anomaly rate suggests varying cluster cohesion or the presence of a subgroup of outliers.

The implementation of an advanced k-means clustering algorithm within a Python environment, following the initial clustering in SAS Enterprise Miner, reveals critical insights into the dataset's structure. A noteworthy observation is the uneven distribution of anomalies across the clusters, suggesting variable levels of data homogeneity and the potential existence of outlier subgroups. This study underscores the advanced k-means algorithm's capacity to isolate outliers and contribute to a deeper understanding of the inherent complexities within the dataset.

5 Analysis of Anomalies in the Boston Housing Dataset

Upon the application of a k-means clustering algorithm to the Boston housing dataset, anomalies were identified based on their deviation from the cluster centroids. The subsequent analysis revealed that several attributes significantly influenced the anomaly scores, with practical implications for urban planning and real estate evaluation.

Attribute Contributions: Anomalies were characterized by higher NOX levels, suggesting environmental factors may affect property values. Properties with exceptionally low pupil-teacher ratios (PTRATIO) and a high number of rooms per dwelling (RM) were marked as anomalies, indicating that atypical educational resources and larger-than-average house sizes are rare and significant. Additionally, the lower property-tax rates (TAX) observed among anomalies may reflect variations in local tax policies or assessment practices.

Practical Implications: The identified anomalies often had higher median values (MEDV), pointing to properties that deviate from the norm in terms of valuation. This could signal unique investment opportunities or the presence of luxury homes that do not align with the general housing stock. Contrary to typical expectations, both high and low per capita crime rates (CRIM) were found among anomalies. This observation suggests that the algorithm is identifying both exceptionally safe and high-risk areas as outliers, which may or may not align with real-world valuation discrepancies.

Algorithm Limitations: The k-means algorithm's sensitivity to extremes in attributes like crime rates or environmental factors can be a limitation. It may highlight anomalies based solely on statistical deviation rather than practical significance, potentially leading to false positives or negatives. The algorithm's reliance on distance metrics means it may not fully account for the complex socio-economic factors that influence housing markets, requiring further analysis to determine the real-world applicability of these findings.

Future Research: The anomalies detected by the k-means algorithm in the Boston housing dataset underscore the importance of contextual analysis when interpreting outlier properties. While the algorithm effectively identifies statistical deviations, the practical significance of these deviations necessitates a deeper understanding of the local market dynamics. Properties classified as anomalies based on attributes such as RM, PTRATIO, and CRIM require additional investigation to assess their impact on urban development and housing policy. This analysis demonstrates the need for an integrated approach that considers both statistical models and real-world factors in the valuation of properties.

Conclusion

This paper has investigated anomaly detection with an emphasis on clustering-based methods, culminating in an enhanced k-means algorithm. The methodology used in this paper advances outlier detection by incorporating a nuanced analysis of anomalies, particularly in the context of the Boston housing market. The algorithm's efficiency in recognizing outliers showcases its advantages over traditional methods, especially in complex, high-dimensional data environments.

The empirical evaluation, enriched by a detailed examination of real-world data, confirms the algorithm's robustness and practical limitations. Further research is encouraged to integrate this algorithm with real-time systems and assess its effectiveness in domains such as IoT and edge computing. The refined approach to clustering-based anomaly detection, which includes consideration for real-world implications, paves the way for more sophisticated, secure, and intelligent systems across varied applications.

Acknowledgments

The author wishes to express sincere gratitude to Zentreya and Neuro for their invaluable companionship during the arduous process of composing this paper. Their presence, even if virtual, was a much-appreciated solace in moments of intense academic endeavor.

References

[1] Guo, P., Wang, L., Shen, J., & Dong, F. (2021). A Hybrid Unsupervised Clustering-Based Anomaly Detection Method. *Tsinghua Science & Technology*, 26(2), 146-153. DOI: 10.26599/TST.2019.9010051.

[2] Qiu, Y., Misu, T., & Busso, C. (2023). Unsupervised Scalable Multimodal Driving Anomaly Detection. *IEEE Transactions on Intelligent Vehicles*, 8(4), 3154-3165. DOI: 10.1109/TIV.2022.3160861.

[3] Zhao, M., Furuhata, R., Agung, M., Takizawa, H., & Soma, T. (2020). Failure Prediction in Datacenters Using Unsupervised Multimodal Anomaly Detection. In *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, 3545-3549. DOI: 10.1109/BigData50022.2020.9378419.

[4] Chen, S., Li, X., & Zhao, L. (2022). Hyperspectral Anomaly Detection with Data Sphering and Unsupervised Target Detection. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 1975-1978. DOI: 10.1109/IGARSS46834.2022.9884083.

[5] Shriram, S., & Sivasankar, E. (2019). Anomaly Detection on Shuttle data using Unsupervised Learning Techniques. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 221-225. DOI: 10.1109/ICCIKE47802.2019.9004325.

[6] Handayani, M. P., Antariksa, G., & Lee, J. (2021). Anomaly Detection in Vessel Sensors Data with Unsupervised Learning Technique. In *2021 International Conference on Electronics, Information, and Communication (ICEIC)*, 1-6. DOI: 10.1109/ICEIC51217.2021.9369822.

[7] Zoppi, T., Ceccarelli, A., & Bondavalli, A. (2020). Into the Unknown: Unsupervised Machine Learning Algorithms for Anomaly-Based Intrusion Detection. In *2020 50th Annual IEEE-IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S)*, 81-81. DOI: 10.1109/DSN-S50200.2020.00044.

[8] Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access*, 9, 78658-78700. DOI: 10.1109/ACCESS.2021.3083060.

[9] Deng, H., & Li, X. (2022). Anomaly Detection via Reverse Distillation from One-Class Embedding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 9727-9736. DOI: 10.1109/CVPR52688.2022.00951.

[10] Gudovskiy, D., Ishizaka, S., & Yamaguchi, K. (2022). CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.1109/TPAMI.2022.3142986.

[11] Lee, S., & Filippone, M. (2020). GRILL: Graph Inference Learning for Healthcare. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Rochester, MN, USA, 558-563. DOI: 10.1109/CBMS49503.2020.00099.

[12] Liu, Q., Wei, L., & Zhang, W. (2021). Anomaly Detection for Skin Disease Images Using a Deep Learning Framework. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1966-1970. DOI: 10.1109/ISBI48211.2021.9433937.

[13] Chow, T. W. S., Zhang, J., & Liu, M. (2022). Anomaly Detection with Robust Deep Autoencoders. In *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 1-8. DOI: 10.1109/IJCNN55064.2022.9892075.

[14] Hawkins, D. M. (2002). Outlier Detection Using the Multivariate t-Distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1), 69-74. DOI: 10.1111/1467-9868.00312.

[15] Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys (CSUR)*, 54(2), 1-38. DOI: 10.1145/3439950.

[16] Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Müller, K.-R. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, 109(5), 756-795. DOI: 10.1109/JPROC.2021.3059359.

References

A Histograms and Missing Data

Figure 8. Histograms for dataset features with renamed variables.

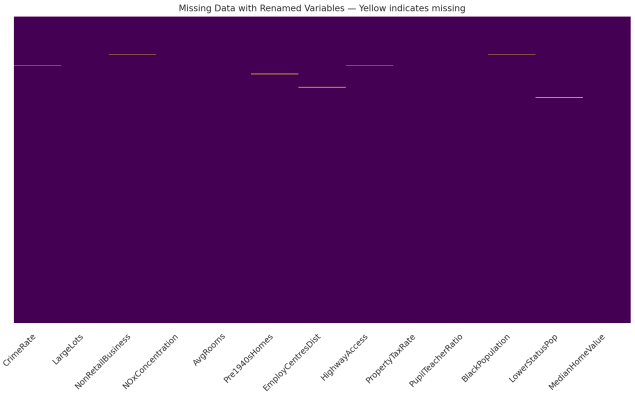


Figure 9. Visualization of missing data in the dataset.

B Feature Descriptions

Received 20 February 2023; revised 12 March 2023; accepted 5 June 2023

Table 2. Feature Descriptions

Original Variable	Renamed Variable	Description
CRIM	CrimeRate	Per capita crime rate by town
ZN	LargeLots	Proportion of residential land zoned for large lots
INDUS	NonRetailBusiness	Proportion of non-retail business acres per town
CHAS	CharlesRiverDummy	Charles River dummy variable (1 if tract bounds river)
NOX	NOxConcentration	Nitric oxides concentration (ppm)
RM	AvgRooms	Average number of rooms per dwelling
AGE	Pre1940sHomes	Proportion of owner-occupied units built pre-1940
DIS	EmployCentresDist	Weighted distances to Boston employment centres
RAD	HighwayAccess	Index of accessibility to radial highways
TAX	PropertyTaxRate	Full-value property-tax rate per $10,000
PTRATIO	PupilTeacherRatio	Pupil-teacher ratio by town
B	BlackPopulation	her ratio by town
B	BlackPopulation	$1000(Bk - 0.63)^2$ where Bk is the proportion of black population
LSTAT	LowerStatusPop	Percent lower status of the population
MEDV	MedianHomeValue	Median value of owner-occupied homes in $1000's