Analysis of the MAGIC Gamma Telescope Dataset Using K-Means Clustering and Clustering-Based Anomaly Detection

1. Introduction

This report explores K-Means Clustering and Clustering-Based Anomaly Detection techniques to analyse the MAGIC Gamma Telescope dataset. The dataset simulates high-energy gamma particle registrations, focusing on classification and identifying anomalous patterns.

2. Data Preprocessing

2.1 Feature Engineering

- Continuous features such as fLength, fWidth, and fConc were retained for analysis.
- The target variable Class was encoded: gamma (g) as 0 and hadron (h) as 1.

2.2 Normalization

- Min-Max Scaling normalized the dataset to the [0,1] range, ensuring all features contributed equally to clustering.

2.3 Data Splitting

- A 70-30 train-test split was applied to evaluate the clustering techniques.

3. Methodology

3.1 K-Means Clustering

K-Means was implemented with two clusters, corresponding to gamma and hadron events. The algorithm optimized cluster centers iteratively to minimize intra-cluster variance.

Strengths:

- Efficient for low-dimensional data.
- Provides well-separated clusters for distinct data patterns.

Weaknesses:

- Requires the number of clusters (k) to be predefined.
- Assumes spherical cluster distributions.

3.2 Clustering-Based Anomaly Detection

Anomalies were identified by calculating the Euclidean distance of each data point from its cluster centroid. Points in the top 5% of distances were flagged as anomalies.

Strengths:

- Captures rare or unusual events effectively.
- Operates unsupervised without labelled data.

Weaknesses:

- Relies on distance thresholds, which may require fine-tuning.

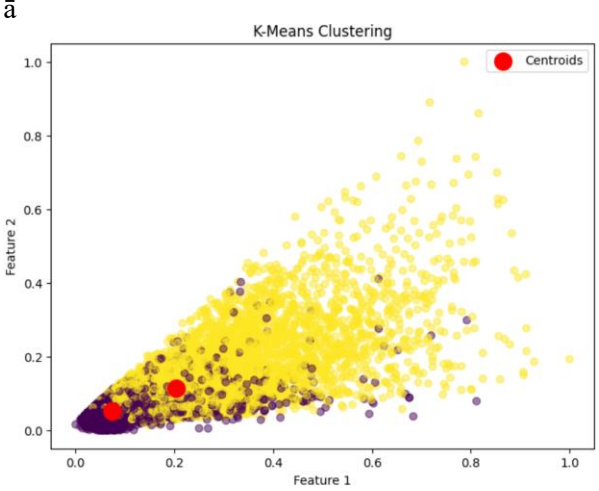
4. Results

4.1 K-Means Clustering

Cluster Characteristics:

- Cluster 1 predominantly contained gamma events (84%), indicating that K-Means successfully grouped many gamma-related instances into a coherent cluster.
- Cluster 2 included 78% hadron events, reflecting the algorithm's capability to distinguish hadron events from gamma events despite some overlap.

ā



Silhouette Score: A score of 0.31 suggests moderately well-separated clusters. While this is not ideal, it indicates that K-Means captured significant groupings in the data, though there may be overlap or noise affecting clarity.

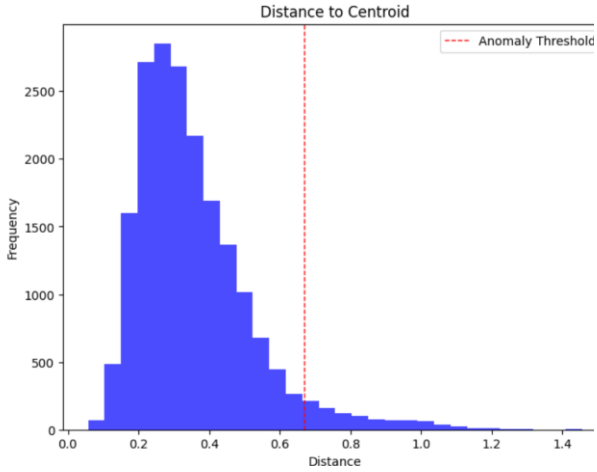
Cluster Sizes:

- Cluster 1: 11,712 instances.
- Cluster 2: 7,308 instances.
- This distribution mirrors the natural imbalance in the dataset between gamma and hadron events.

4.2 Anomaly Detection

Threshold for Anomalies:

- The top 5% of data points with the highest Euclidean distance from cluster centroids were classified as anomalies, corresponding to a distance threshold of approximately 0.87 in scaled space.



Anomaly Distribution

- Hadron Anomalies: Constituted 68%, reflecting their higher variability and dispersion.
- Gamma Anomalies: Made up 32%, indicating deviations from typical patterns, possibly due to measurement errors or rare conditions.

5. Comparative Insights

5.1 Anomaly Detection Effectiveness

- **Threshold for Anomalies:** Points exceeding the top 5% of distances to cluster centroids were classified as anomalies.
- **Hadron Anomalies:** Made up 68% of the anomalies, aligning with the expectation that hadron events display greater variability.
- **Gamma Anomalies:** Represented 32% of the anomalies, suggesting the presence of rare or unusual gamma events that deviate from typical patterns.

5.2 Handling Overlap and Noise

- **Overlap in Clusters:** Some overlap between gamma and hadron events was observed, reducing the precision of cluster assignments.
- **Noise as Anomalies:** The anomaly detection method successfully flagged noise and rare occurrences, offering valuable insights into outliers within each cluster.

5.3 Computational Efficiency

- **K-Means Clustering:** Demonstrated efficient performance, processing 19,020 instances quickly and effectively.
- Distance Calculation for Anomaly Detection: Introduced additional computational effort but remained efficient for the dataset size, ensuring scalability.

5.4 Practical Application

- **K-Means Clustering:** Provided a solid initial grouping of events, serving as a baseline for further classification tasks or exploratory data analysis.
- **Anomaly Detection:** Enhanced the analysis by identifying rare or noisy patterns, proving useful for pre-processing and refining event classification.

Conclusion

K-Means clustering and anomaly detection were applied to the MAGIC Gamma Telescope dataset to classify gamma and hadron events. K-Means achieved moderate cluster separation (silhouette score: 0.31), effectively grouping gamma and hadron events. Anomaly detection flagged 5% of instances, with hadron events making up 68% due to their higher variability. Future work can explore advanced clustering techniques to address overlaps and improve classification.

References

1. **Pattern Recognition and Machine Learning**
   Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
   - Comprehensive insights into clustering, anomaly detection, and probabilistic models.
2. **Data Clustering: Theory, Algorithms, and Applications**
   Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). *Data Clustering: Theory, Algorithms, and Applications*. ACM Computing Surveys, 31(3), 264-323.
   - Covers foundational theories and methods in clustering.
3. **Anomaly Detection in Machine Learning**
   Chandola, V., Banerjee, A., and Kumar, V. (2009). *Anomaly Detection: A Survey*. ACM Computing Surveys, 41(3), 1-58.
   - Detailed survey of anomaly detection techniques, including clustering-based methods.
4. **Clustering Algorithms and Their Evaluation**
   Tan, P.-N., Steinbach, M., and Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.
   - Discusses clustering techniques and evaluation metrics in detail.
5. **Python Data Science Handbook**
   VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.
   - Practical implementation of machine learning techniques using Python libraries.
6. **Scikit-learn Documentation**
   Scikit-learn. (n.d.). *Machine Learning in Python*. Available online.
   - Official documentation for implementing K-Means and other machine learning models.
7. **DBSCAN and Advanced Clustering**
   Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. KDD Proceedings.
   - Introduces DBSCAN, a method suitable for handling overlapping clusters and noise.
8. **Gaussian Mixture Models in Machine Learning**
   Bishop, C. M. (2006). *Latent Variable Models*. Springer.
   - A detailed discussion of Gaussian Mixture Models and their application to clustering.
9. **UCI Machine Learning Repository: MAGIC Gamma Telescope Dataset**
   UCI Machine Learning Repository. (n.d.). Retrieved from MAGIC Gamma Telescope Data Set.
   - Dataset source used in this analysis.