Network Traffic Profiling - K-Means Clustering

Unsupervised Learning

---

1. Introduction

You are a network analyst at a large organization. The IT department has collected network flow data but lacks labeled information about what types of traffic are present. Your manager needs to understand the different patterns of network usage to optimize network resources, identify normal behavior baselines, and improve security monitoring.

Important: This is unlabeled network traffic data. Unlike supervised learning, you don't have predefined categories. Your task is to discover natural groupings in the data using clustering techniques.

---

2. Business Objective

The organization wants to:

1. Identify distinct patterns of network traffic behavior
2. Profile different types of network usage (web, streaming, file transfer, etc.)
3. Understand resource utilization by traffic type
4. Create baselines for anomaly detection
5. Optimize network policies and QoS (Quality of Service) settings

Your Task:

1. Clean and prepare the unlabeled network traffic dataset
2. Use K-Means clustering to discover natural groupings
3. Determine the optimal number of clusters
4. Profile and interpret each cluster
5. Provide actionable recommendations for network management

---

3. Dataset Description

File: `network_traffic_raw.csv`
Size: Approximately 1,800+ network flow records
Type: Unlabeled data (unsupervised learning)

Features:

Feature	Type	Description
FlowID	Numeric	Unique identifier for each flow
FlowDuration	Numeric	Duration of the flow in seconds
TotalPackets	Numeric	Total number of packets in the flow
TotalBytes	Numeric	Total bytes transmitted in the flow
AvgPacketSize	Numeric	Average packet size in bytes
BytesPerPacket	Numeric	Calculated: TotalBytes / TotalPackets
PacketRate	Numeric	Packets per second
ByteRate	Numeric	Bytes per second (throughput)
FlowVolume	Numeric	Total data volume: Duration × ByteRate
Port	Numeric	Destination port number
Protocol	Categorical	Network protocol (TCP, UDP, etc.)
FIN_Flag_Count	Numeric	Number of FIN flags
SYN_Flag_Count	Numeric	Number of SYN flags
RST_Flag_Count	Numeric	Number of RST flags
PSH_Flag_Count	Numeric	Number of PSH flags
ACK_Flag_Count	Numeric	Number of ACK flags
InterArrivalTime	Numeric	Average time between packets (seconds)

---

4. Your Tasks

Task 1: Data Quality Assessment

- Load and explore the dataset
- Identify ALL data quality issues
- Document issues with statistics and examples

- Create visualizations showing data problems
- Write a data quality report

Task 2: Data Cleaning and Preprocessing

- Clean the dataset systematically
- Handle missing values appropriately
- Remove or correct outliers and invalid values
- Standardize categorical variables
- Fix logical inconsistencies between calculated fields
- Document all decisions with justifications
- Save cleaned dataset as `network_traffic_clean.csv`

Task 3: Exploratory Data Analysis

- Analyze feature distributions
- Identify correlations between features
- Look for potential groupings visually
- Create meaningful visualizations
- Document patterns observed

Task 4: Feature Engineering and Scaling

- **CRITICAL:** Feature scaling is MANDATORY for K-Means
- Select relevant features for clustering
- Standardize or normalize features appropriately
- Explain your scaling choice (StandardScaler vs MinMaxScaler)
- Consider creating derived features if useful
- Handle categorical variables (encode or exclude from clustering)

Task 5: Determining Optimal K

- Use multiple methods to find optimal number of clusters:
    - **Elbow Method** (Within-Cluster Sum of Squares - WCSS)
    - **Silhouette Score**
    - **Gap Statistic** (optional but recommended)
- Test range of K values (e.g., K=2 to K=10)
- Create visualizations for each method
- Justify your final choice of K
- Discuss trade-offs between different K values

Task 6: K-Means Clustering

- Implement K-Means with optimal K
- Set random_state for reproducibility
- Consider running multiple initializations (n_init parameter)
- Assign cluster labels to all data points
- Save clustered data with labels

Task 7: Cluster Evaluation

- Calculate clustering quality metrics:
    - Silhouette Score (overall and per-cluster)
    - Davies-Bouldin Index
    - Calinski-Harabasz Index
- Interpret what these metrics mean
- Assess cluster separation and cohesion

Task 8: Cluster Profiling and Interpretation

- Analyze characteristics of each cluster:
    - Statistical summary (mean, median, std) for each feature
    - Dominant protocols and ports
    - Typical flow characteristics
- Name/label each cluster based on behavior patterns
- Create visualization comparing clusters
- Identify what makes each cluster unique

Task 9: Business Insights and Recommendations

- Interpret clusters in business context
- Provide specific recommendations for:
    - Network resource allocation
    - Quality of Service (QoS) policies
    - Security monitoring priorities
    - Bandwidth management
- Discuss how findings can be used operationally

---

5. Important Considerations

K-Means Requirements:

- **Feature Scaling:** K-Means uses Euclidean distance, so features must be on similar scales
- **Numerical Features:** K-Means works with numerical features only
- **Initialization:** Results can vary with random initialization (use random_state)
- **K Selection:** The number of clusters is not known beforehand
- **Outliers:** K-Means is sensitive to outliers (clean data well!)

Network Traffic Context:

- **Common traffic types:** Web browsing, video streaming, file transfers, DNS queries, gaming, IoT
- **Port significance:**

    - 80/443: HTTP/HTTPS (web)
    - 53: DNS
    - 21/22: FTP/SSH (file transfer)
    - 1883/8883: MQTT (IoT)
    - High ports (>1024): Often gaming or P2P
- **Behavior patterns:**
    - Streaming: High duration, large bytes, consistent rate
    - DNS: Very short, few packets, UDP
    - Web: Moderate size, bursty, HTTP/HTTPS
    - File transfer: Large bytes, high rate, TCP

Expected Challenges:

- Some flows may not fit neatly into clusters
- Optimal K is subjective and context-dependent
- Different features may suggest different clusters
- Outliers can distort cluster centroids

---

6. Deliverables

1. Jupyter Notebook (.ipynb):

- **Part A:** Data loading and quality assessment
- **Part B:** Data cleaning (well-documented)
- **Part C:** Exploratory data analysis
- **Part D:** Feature engineering and scaling
- **Part E:** Optimal K determination (multiple methods)
- **Part F:** K-Means clustering implementation
- **Part G:** Cluster evaluation
- **Part H:** Cluster profiling and interpretation
- **Part I:** Visualizations and insights

2. Cleaned Dataset:

- `network_traffic_clean.csv` - Cleaned data ready for clustering

3. Presentation:

- **Section 1:** Data quality issues and cleaning approach
- **Section 2:** Feature selection and scaling decisions
- **Section 3:** Optimal K determination process
- **Section 4:** Cluster analysis and profiling
    - Description of each cluster
    - Statistical comparison
    - Business interpretation

- **Section 5:** Recommendations for network management
- **Section 6:** Limitations and future work

4. Required Visualizations:

- Data quality issues charts
- Feature distributions and correlations
- Elbow curve (WCSS vs K)
- Silhouette score plot
- 2D cluster visualization (using PCA or t-SNE for dimensionality reduction)
- Cluster profiles (parallel coordinates plot or heatmap)
- Cluster size distribution

---

7. K-Means Specific Guidance

Why Feature Scaling Matters:

```
# Example: Without scaling
Feature1: [0.1, 0.2, 0.3]  # Small range
Feature2: [1000, 2000, 3000]  # Large range
# K-Means will be dominated by Feature2!

# With scaling (StandardScaler):
Feature1: [-1.22, 0, 1.22]
Feature2: [-1.22, 0, 1.22]
# Now both features contribute equally
```

Elbow Method:

- Plot WCSS (Within-Cluster Sum of Squares) vs K
- Look for "elbow" where improvement slows
- Not always clear-cut!

Silhouette Score:

- Range: -1 to +1
- Values close to +1: Well-clustered
- Values close to 0: On cluster boundary
- Negative values: Possibly wrong cluster
- Higher average = better clustering

Cluster Interpretation Tips:

- Compare cluster centroids
- Look at feature means per cluster

- Identify dominant ports/protocols per cluster
- Visualize in 2D using PCA or t-SNE
- Give meaningful names based on behavior

---

8. Submission Guidelines

- Filename: `TestAA_TrafficClustering.zip`
- Include: Notebook, cleaned CSV, Presentation
- Ensure notebook runs without errors
- Use markdown cells to explain your work

---

9. Hints (Strategic Guidance)

Getting Started:

- Clean data thoroughly BEFORE clustering
- Outliers will severely impact K-Means results
- Not all features may be useful for clustering
- Scale features BEFORE running K-Means
- Try different random_state values to ensure stability

Choosing K:

- No single "correct" answer
- Consider multiple evaluation methods
- Business context matters (too many clusters = not useful)
- Typical range for network traffic: 3-7 clusters
- Balance between granularity and interpretability

Common Mistakes to Avoid:

- Forgetting to scale features (CRITICAL ERROR!)
- Including FlowID in clustering
- Not handling missing values before clustering
- Choosing K without justification
- Not interpreting what clusters mean
- Ignoring outliers
- Not considering domain knowledge

Success Indicators:

- Clusters have clear, interpretable differences
- Silhouette scores are positive and reasonable (>0.3)

- Elbow curve shows clear structure
- Cluster assignments make sense given features
- Business recommendations are specific and actionable

---

10. Expected Outcomes

After proper implementation, you should discover:

- Distinct traffic patterns (e.g., bulk transfer vs interactive)
- Resource-intensive vs lightweight traffic
- Different protocol behaviors
- Temporal patterns (long vs short flows)
- Application-specific characteristics

Your clusters should be:

- Separable (distinct from each other)
- Cohesive (similar within cluster)
- Interpretable (can explain what they represent)
- Actionable (can guide business decisions)

Good luck! Unsupervised learning requires creativity in interpretation. Focus on discovering meaningful patterns and telling a compelling story about what you find in the data.