Advanced Transportation Mode Classification: A Multi-Stage Machine Learning Approach for Detecting Hidden Mobility Patterns

Abstract

Transportation mode classification represents a critical challenge in mobility analytics, particularly when addressing scenarios where travel patterns are intentionally obscured (hiding behavior). This research presents a comprehensive machine learning pipeline that combines statistical analysis, advanced feature engineering, and ensemble methods to achieve 93% accuracy in classifying walking, biking, and driving behaviors. Our approach integrates Interquartile Range (IQR) analysis, Principal Component Analysis (PCA), Density-Based Spatial Clustering (DBSCAN), and Gaussian Mixture Models (GMM) to create a robust foundation for ensemble classification. The methodology proves especially effective in detecting deceptive movement patterns, making it well-suited for applications requiring accurate mode detection under adversarial conditions.

Keywords: Transportation mode classification, ensemble learning, mobility analytics, unsupervised learning, feature engineering

1. Introduction

1.1 Problem Statement

In the context of transportation behavior analysis, determining the mode of travel for each trip presents significant challenges, especially when individuals may deliberately obscure or vary their travel modes (hiding behavior). While mode choice seeking algorithms are of lower priority, accurately identifying transportation modes during hiding scenarios is of critical importance. The core challenge lies in correctly classifying travel modes (walking, biking, driving) without explicit labels, particularly when travel patterns are intentionally obfuscated.

1.2 Research Objectives

This research aims to:

- Develop a robust transportation mode classification system resistant to hiding behaviors
- Integrate statistical and machine learning approaches for comprehensive pattern detection
- Achieve high accuracy across multiple transportation modes
- Provide interpretable insights into feature importance and model decision-making
- Create a scalable pipeline applicable to diverse mobility datasets

1.3 Contributions

Our key contributions include:

1. A novel multi-stage pipeline combining statistical thresholding with advanced ML techniques
2. Comprehensive feature engineering incorporating temporal, spatial, and contextual factors
3. Integration of unsupervised learning (GMM) to enhance supervised classification
4. Robust outlier detection and dimensionality reduction strategies
5. Achieved 93% overall accuracy with balanced performance across all transportation modes

2. Methodology

2.1 Dataset Overview

The analysis utilizes a comprehensive transportation dataset containing 44,753 trips with the following key attributes:

- **Spatial features**: Start/end coordinates, distance, origin-destination pairs
- **Temporal features**: Trip duration, start time, day of week
- **Derived metrics**: Speed, route density, directional flow patterns
- **Contextual variables**: Rush hour indicators, weekend flags, time periods

2.2 Pipeline Architecture

Our methodology follows a five-stage progressive refinement approach (see Figure 1: Methodology Flowchart):

```
Raw Data → Statistical Analysis → Feature Engineering →
Preprocessing → Unsupervised Discovery → Ensemble Classification
```

The pipeline is designed to incrementally build classification capability, starting with explainable statistical foundations and progressively adding sophisticated machine learning components. Each stage contributes measurable improvement to the final accuracy, as demonstrated in our performance evolution analysis.

3. Stage 1: Statistical Foundation with IQR Analysis

3.1 Interquartile Range Methodology

We employed IQR analysis to establish data-driven thresholds for initial mode categorization, avoiding arbitrary hard-coded cutoffs. This approach offers several advantages:

- **Adaptive thresholds**: Calculations based directly on dataset characteristics
- **Outlier resilience**: Uses quartiles (Q1, Q3) less affected by extreme values
- **Statistical rigor**: Based on established statistical principles rather than assumptions
- **Transparency**: Provides explainable baseline for complex model comparison

3.2 IQR Results and Thresholds

Speed Analysis:

- Q1 = 5.51 km/h, Q3 = 37.90 km/h, IQR = 32.39 km/h
- Thresholds: Walk ≤ 5.51 km/h, Bike ≤ 37.90 km/h, Car > 37.90 km/h

Distance Analysis:

- Q1 = 0.93 km, Q3 = 3.24 km, IQR = 2.31 km
- Thresholds: Walk ≤ 0.93 km, Bike ≤ 3.24 km, Car > 3.24 km

3.3 Initial Classification Performance

The IQR-based classification revealed distinct mode signatures:

- **Walking**: Mean speed 3.06 km/h, mean distance 0.50 km
- **Biking**: Mean speed 12.73 km/h, mean distance 1.80 km
- **Driving**: Mean speed 45.20 km/h, mean distance 4.57 km

4. Stage 2: Advanced Feature Engineering

4.1 Origin-Destination (OD) Pair Analysis

We developed sophisticated metrics to capture route popularity and directional flow (detailed in Figure 3):

Core OD Features:

- `od_pair_density`: Frequency of trips sharing start-end points
- `reverse_od_pair_density`: Trips in opposite direction
- `od_direction_ratio`: Balance measure (od_pair_density / reverse_od_pair_density)

Route Characteristics:

- `is_balanced_route`: Routes with od_direction_ratio between 0.5-2.0
- `is_one_way_route`: Routes with no return trips
- `speed_density_ratio`: Speed normalized by route popularity

4.2 Temporal Context Features

Time-based Engineering:

- Rush hour indicators (`morning_rush`, `evening_rush`)
- Weekend vs. weekday classification

- Hour-of-day and day-of-week encoding
- Time period categorization (6 distinct periods)

Rush Hour Optimization (Figure 3): Through iterative threshold testing using decision tree feature importance analysis, we optimized rush hour detection:

- **Car rush hour**: Distance > 1.5 km, Speed > 35 km/h (84.46% accuracy)
- **Bike rush hour**: Distance 0.3-3.0 km, Speed 3-25 km/h

4.3 Feature Correlation Analysis

Key relationships identified through correlation matrix analysis (Figure 3):

- **Distance-Speed correlation (0.73)**: Longer trips associated with higher speeds
- **Speed-Duration negative correlation (-0.61)**: Faster modes reduce travel time
- **Density-Direction correlation (0.52)**: Popular routes show balanced flow

These correlations validated our feature engineering approach and guided PCA component interpretation.

5. Stage 3: Preprocessing and Dimensionality Reduction

5.1 Principal Component Analysis (PCA)

Configuration:

- Retained 95% of variance using 11 components (reduced from 20 features)
- Applied StandardScaler for feature normalization
- One-hot encoded categorical variables

Benefits Achieved:

- Noise reduction through variance-based selection
- Multicollinearity elimination between correlated features
- Computational efficiency improvement
- Overfitting prevention

5.2 Outlier Detection with DBSCAN

Methodology:

- Used k-distance plot analysis for optimal epsilon selection
- Applied KneeLocator algorithm for objective threshold determination
- Optimal epsilon = 2.3002 based on 5th nearest neighbor distances

Results:

- Identified 12 outliers from 44,753 trips (0.027%)
- Outliers characterized by: higher speeds (~42 vs 19 km/h), longer distances (~5.4 vs 2.3 km)
- Most outliers represented atypical car trips or ambiguous mode boundaries

6. Stage 4: Unsupervised Pattern Discovery

6.1 Gaussian Mixture Model Implementation

Model Selection (Figure 3: GMM Analysis):

- Used Bayesian Information Criterion (BIC) for optimal component selection
- Selected 7 components based on minimum BIC score (Figure 3 shows BIC optimization curve)
- Applied full covariance type with regularization (reg_covar=1e-4)

Cluster Analysis:

- **Cluster 2**: 69.3% bikes, 30.7% walking (primarily biking)
- **Cluster 5**: 63.7% cars, 35.8% bikes (primarily driving)
- **Mixed clusters**: Revealed transitional and ambiguous travel patterns

The GMM analysis (Figure 3) reveals that transportation behavior exists along a continuum rather than discrete categories. This probabilistic understanding proves crucial for handling edge cases and detecting deceptive travel patterns.

6.2 GMM Feature Enhancement

The GMM generated 7 probability features for each trip, capturing:

- **Mode-specific patterns**: Components 1, 5, 6 strongly predict bikes, walking, cars respectively
- **Transitional behaviors**: Components 0, 3, 4 show mixed characteristics
- **Uncertainty quantification**: Enables handling of ambiguous cases

Impact on Feature Space: The addition of GMM probabilities enriched our feature space from 11 PCA components to 18 total dimensions, providing the ensemble model with nuanced pattern information that significantly improved classification accuracy (from 91% to 92% before final ensemble).

7. Stage 5: Ensemble Model Development

7.1 Model Architecture

Base Models:

- **Random Forest**: Optimized through GridSearchCV
  - n_estimators = 200

    - max_features = 'sqrt'
    - No depth restrictions for pattern flexibility
- **Gradient Boosting**: Focused on edge case learning

Ensemble Strategy:

- Soft voting combining predicted probabilities
- Final feature space: 11 PCA components + 7 GMM probabilities (18 dimensions)

7.2 Hyperparameter Optimization

Grid search parameters for Random Forest:

```python
{
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2']
}
```

8. Results and Performance Analysis

8.1 Classification Performance

Overall Metrics:

- **Accuracy**: 93%
- **Weighted F1-Score**: 0.92
- **Macro F1-Score**: 0.92

Per-Class Performance (see Figure 2: Performance Analysis):

Mode	Precision	Recall	F1-Score	Accuracy
Car	94%	97%	96%	97%
Bike	91%	95%	93%	94%
Walk	95%	84%	89%	84%

8.2 Confusion Matrix Analysis

The confusion matrix reveals important insights into model behavior (Figure 2):

Key Insights:

- Cars rarely misclassified (only 127 car trips misclassified as bikes)
- Walking-biking confusion most common (582 walking trips misclassified as bikes)
- No confusion between cars and walking (zero misclassifications in both directions)
- Misclassification pattern follows logical mode similarity hierarchy

This confusion pattern validates our understanding of transportation mode relationships, where walking and biking share more characteristics (speed ranges, infrastructure usage) than either mode does with driving.

8.3 Feature Importance Analysis

Top 5 Most Important Features (Figure 2):

1. **PCA Component 0 (20.1%)**: Speed and efficiency patterns
   - Positive: speed_density_ratio (+0.464), speed_kmh (+0.426)
   - Negative: od_pair_density (-0.333), trip_duration (-0.313)
2. **PCA Component 7 (19.0%)**: Rush hour context
   - Positive: rush_hour_car_feature (+0.739)
   - Negative: rush_hour_bike_feature (-0.415), distance (-0.405)
3. **PCA Component 1 (11.7%)**: Route characteristics
   - Positive: od_pair_density (+0.432), od_direction_ratio (+0.331)
   - Negative: is_balanced_route (-0.341)
4. **PCA Component 10 (9.0%)**: Trip efficiency
   - Positive: speed_density_ratio (+0.743)
   - Negative: distance (-0.432)
5. **PCA Component 5 (5.6%)**: Temporal patterns
   - Positive: trip_duration (+0.654), is_one_way_route (+0.445)

8.4 Model Evolution and Accuracy Progression

Figure 2 illustrates the progressive improvement achieved through each pipeline stage:

- **IQR Baseline**: ~85% accuracy
- **+ Feature Engineering**: ~88% accuracy
- **+ PCA**: ~90% accuracy
- **+ DBSCAN**: ~91% accuracy
- **+ GMM**: ~92% accuracy
- **Final Ensemble**: 93% accuracy

This demonstrates the value of our multi-stage approach, where each component contributes measurable improvement while maintaining interpretability.

9. Key Insights and Implications

9.1 Speed-Distance Relationship as Core Signal

The consistent emergence of speed-distance interactions across all analysis stages confirms this relationship as the fundamental discriminator for transportation mode classification. This validates the biological and mechanical constraints that naturally separate walking, biking, and driving behaviors.

9.2 Contextual Features Provide Critical Nuance

Beyond core metrics, features such as OD pair density, directional flow, and rush hour dynamics provide essential context. These enable the model to understand not just *how* people travel, but *why* certain modes are chosen under specific conditions.

9.3 Probabilistic Modeling Captures Uncertainty

The GMM integration revealed that transportation behavior exists along a spectrum rather than discrete categories. This probabilistic approach proves especially valuable for handling ambiguous cases and detecting deceptive travel patterns.

9.4 Rush Hour Patterns Drive Mode Choice

The high importance of rush hour indicators (PCA Component 7) demonstrates that temporal context significantly influences transportation decisions. Peak-hour behaviors exhibit distinctive characteristics that require separate modeling consideration.

10. Applications and Future Work

10.1 Practical Applications

Urban Planning:

- Traffic flow optimization based on mode-specific patterns
- Infrastructure development guided by actual usage patterns
- Public transit route planning using mobility insights

Security and Surveillance:

- Detection of anomalous travel patterns
- Identification of deceptive movement behaviors
- Enhanced location-based security systems

Transportation Services:

- Dynamic routing optimization for ride-sharing platforms
- Predictive maintenance scheduling based on usage patterns
- Personalized transportation recommendations

10.2 Future Research Directions

1. **Real-time Classification**: Adaptation for streaming data processing
2. **Multi-modal Trips**: Extension to handle combined transportation modes
3. **Environmental Integration**: Incorporation of weather, traffic, and event data
4. **Scalability Testing**: Validation across different geographic regions and populations
5. **Adversarial Robustness**: Enhanced detection of intentional pattern obfuscation

11. Conclusion

This research demonstrates that sophisticated transportation mode classification can be achieved through a carefully designed multi-stage pipeline combining statistical foundations, advanced feature engineering, and ensemble machine learning. Our approach achieves 93% accuracy while maintaining interpretability and robustness against deceptive travel patterns.

The key innovation lies in the progressive refinement strategy: beginning with explainable statistical thresholds (IQR analysis), enriching the feature space through contextual engineering (OD pair analysis, rush hour patterns), leveraging unsupervised learning for pattern discovery (GMM clustering), and culminating in an optimized ensemble classifier. This methodology proves particularly effective for detecting hidden or adversarial mobility behaviors, making it valuable for applications requiring robust mode detection under challenging conditions.

The integration of spatial, temporal, and behavioral signals creates a comprehensive understanding of human mobility patterns that extends beyond simple speed-distance relationships. The probabilistic components provided by GMM enable nuanced handling of uncertain cases, while the ensemble approach ensures robust performance across diverse travel scenarios.

Technical Contributions:

1. **Multi-stage progressive refinement**: Each stage contributes measurable improvement (85% → 93% accuracy)
2. **Feature engineering innovations**: OD pair density analysis, rush hour context, speed-density ratios
3. **Unsupervised enhancement**: GMM probabilities as features improve ensemble performance
4. **Robust preprocessing**: PCA dimensionality reduction and DBSCAN outlier detection
5. **Ensemble optimization**: Soft voting combination of Random Forest and Gradient Boosting

Performance Validation: Our confusion matrix analysis shows the model successfully distinguishes between transportation modes with minimal cross-contamination between cars and walking (zero

misclassifications), while handling the natural overlap between walking and biking appropriately. The feature importance analysis confirms that our engineered features (particularly speed-efficiency patterns and rush hour context) drive classification decisions.

This work establishes a new benchmark for transportation mode classification in scenarios involving potential hiding behavior, with applications ranging from urban planning to security analytics.

List of Figures

- **Figure 1**: Transportation Mode Classification Pipeline - Complete methodology flowchart showing the five-stage progressive refinement approach
- **Figure 2**: Performance Analysis and Results - Confusion matrix, per-class metrics, feature importance ranking, and accuracy evolution timeline
- **Figure 3**: Feature Engineering and GMM Analysis - Detailed view of feature correlation analysis, GMM clustering results, and rush hour pattern optimization

References

1. Van Goeverden, Cornelis Dirk. "The Value of Travel Speed." *Transportation Research Interdisciplinary Perspectives*, vol. 13, Mar. 2022, p. 100530.
2. Zhou, Zewei, et al. "A Comprehensive Study of Speed Prediction in Transportation System: From Vehicle to Traffic." *iScience*, vol. 25, no. 3, Mar. 2022, p. 103909.
3. Tang, Jinjun, et al. "Exploring Urban Travel Patterns Using Density-Based Clustering with Multi-Attributes from Large-Scaled Vehicle Trajectories." *Physica A: Statistical Mechanics and Its Applications*, vol. 561, Jan. 2021, p. 125301.
4. Pitombo, Cira Souza, and Monique Martins Gomes. "Study of Work-Travel Related Behavior Using Principal Component Analysis." *Open Journal of Statistics*, vol. 4, no. 11, Dec. 2014, pp. 889-901.

Corresponding Author: [Your Name]
Affiliation: [Your Institution]
Email: [Your Email]