Synthetic Anomaly Generation Methods for Isolation Forest Evaluation

Overview

This report details three different methods implemented for generating synthetic anomalies to evaluate Isolation Forest performance: Gaussian, Uniform, and Extreme value methods. Each method is designed to simulate different types of anomalies that might occur in real-world scenarios.

Data Preparation

Before generating synthetic anomalies, we extract key statistical properties from the normal data: - Mean ( ): Central tendency of each feature - Standard Deviation ( ): Spread of each feature - Minimum values: Lower bounds of normal data - Maximum values: Upper bounds of normal data

Method 1: Gaussian Distribution

Implementation

```
synthetic_anomalies = np.random.normal(mean, 3 * std, size=(n_anomalies, data.shape[1]))
```

Methodology

- Uses a normal distribution with the same mean as the original data
- Standard deviation is tripled (3) to generate values outside normal range
- Approximately 99.7% of normal data falls within ±3 of the mean
- Generated anomalies are likely to fall outside this range

Characteristics

- Preserves feature correlations
- Generates both moderate and extreme anomalies
- More concentrated around the boundaries of normal data
- Suitable for detecting gradual anomalies

Method 2: Uniform Distribution

Implementation

```
synthetic_anomalies = np.random.uniform(
    low=min_vals - 2 * std,
    high=max_vals + 2 * std,
    size=(n_anomalies, data.shape[1])
)
```

Methodology

- Generates values uniformly between expanded boundaries
- Lower bound: minimum value - 2
- Upper bound: maximum value + 2
- Creates a wider range than the original data

Characteristics

- Equal probability across the expanded range
- Does not preserve feature correlations
- Good for testing boundary detection
- Simulates random, unexpected values

Method 3: Extreme Value

Implementation

```
multipliers = np.random.choice([-3, 3], size=(n_anomalies, data.shape[1]))
synthetic_anomalies = mean + multipliers * std
```

Methodology

- Generates values exactly at ± 3 from the mean
- Uses random choice between positive and negative extremes
- Creates clear outliers at specific distances

Characteristics

- Produces distinct anomalies
- Always generates significant deviations
- Good for testing extreme value detection
- Simulates systematic errors or critical events

Comparative Analysis

Strengths and Use Cases:

1. Gaussian Method:
   - Best for: Gradual anomaly detection
   - Natural deviation patterns
   - Maintains data relationships
2. Uniform Method:
   - Best for: Boundary testing
   - Wide coverage of possible values
   - Tests model flexibility
3. Extreme Method:
   - Best for: Critical anomaly detection

- Clear separation from normal data
- Tests sensitivity thresholds

Selection Criteria

Choose the method based on: - Expected real-world anomaly patterns - Desired sensitivity testing - Nature of the monitored system

Implementation Considerations

Data Dimensionality

- All methods handle multi-dimensional data
- Feature-wise statistics preserve data structure
- Correlations handled differently by each method

Scale Independence

- Methods adapt to data scale through statistics
- No need for pre-scaling of features
- Maintains relative importance of features

Usage Recommendations

1. For System Testing:
   - Use all three methods to ensure robust detection
   - Compare performance across methods
   - Identify potential blind spots
2. For Production:
   - Select method matching expected anomalies
   - Consider combining methods
   - Adjust parameters based on domain knowledge
3. For Model Tuning:
   - Use methods sequentially
   - Start with extreme values
   - Gradually test with more subtle anomalies