KATHMANDU UNIVERSITY

Department of Artificial Intelligence

Dhulikhel, Kavre

Lab Report On: Applications and Case Studies

Submitted By:

Name: Aaryan Shakya

Roll No: 20

Subject Code: AICC 301

Submitted To:

Mr. Sunil Regmi

Lecturer, Department of Artificial Intelligence

Submission Date:

6th July 2025

Objectives

1. Understand the concept and types of outlier detection in data mining.
2. Learn various methods used for detecting outliers in different datasets.
3. Explore real-world applications of outlier detection in diverse domains such as education, business, spatial analysis, and text mining.

Introduction

Outlier Detection

Outliers are data points that deviate significantly from the majority of observations and can indicate important, sometimes critical, information. Recognizing the different types helps tailor detection techniques appropriately.

Types of Outliers

- **Point Outliers:** These are individual anomalous values that stand apart from the rest.
  Example: In credit card fraud detection, a single transaction of an unusually high amount compared to a customer's typical spending pattern is a point outlier. Detecting such a transaction quickly can prevent financial loss.

- **Contextual Outliers:** These depend on the context or environment. The same value may be normal in one context but abnormal in another.
  Example: In climate monitoring, a temperature of 40°C is normal in summer but is a contextual outlier in winter. Detecting this helps in climate anomaly detection and alerts.

- **Collective Outliers:** These are groups of data points that collectively deviate from the norm, even if individual points may not be outliers alone.
  Example: In network security, a sudden burst of data packets coming from a specific IP address may indicate a Distributed Denial of Service (DDoS) attack. Individually, packets may seem normal, but collectively they form a collective outlier signaling an attack.

Methods of Outlier Detection

Different datasets and problem contexts require different methods for effective outlier detection.

- **Z-score Method:**
  This method standardizes data by measuring how far each point is from the average in terms of standard deviations.
  Real-world use: Detecting unusual sensor readings in manufacturing equipment that deviate sharply from normal operating conditions. For example, if a temperature sensor usually reads around 70°F, a sudden reading of 150°F would have a high Z-score and be flagged.

- **Interquartile Range (IQR):**
  Focuses on values that lie far outside the typical range between the 25th and 75th percentiles.
  Real-world use: In financial datasets, it can identify anomalous stock price changes that fall outside normal volatility, helping traders detect potential market manipulation.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**
  Useful for spatial or geographical data where clusters of data points are expected, and anomalies lie outside these clusters.
  Real-world use: Environmental monitoring systems use DBSCAN to detect illegal logging by finding clusters of normal forest activity and identifying unusual deforestation patterns as outliers.

- **Isolation Forest:**
  An efficient algorithm that isolates anomalies by partitioning data. Points that are easier to isolate are more likely outliers.
  Real-world use: Used by cybersecurity firms to detect rare, suspicious behavior in user activity logs, like an account suddenly logging in from a foreign country or accessing unusual resources.

Applications of Outlier Detection with Real-World Examples

a. Educational Data Mining

- **Use case:** Universities track student performance and engagement. Outlier detection can spot students whose sudden drop in attendance or grades may indicate academic struggles or disengagement.

- **Example:** A student who consistently scores above average but suddenly fails multiple exams may be flagged for early counseling or tutoring interventions.

b. Business Intelligence

- **Use case:** Retailers analyze sales data to identify fraudulent transactions or unusual customer behavior.

- **Example:** An online retailer detects outliers in transaction patterns, such as multiple high-value purchases from a new account in a short time frame, which may signal fraud.

c. Spatial Data Mining

- **Use case:** Government agencies monitor geographic data for anomalies indicating illegal activities or natural disasters.

- **Example:** Satellite imagery processed with clustering algorithms identifies unusual deforestation patches in protected forest areas, triggering inspections.

d. Time Series and Text Mining

- **Time Series Example:** Stock markets use outlier detection to identify abnormal price movements or trading volumes that could indicate market manipulation or insider trading.
- **Text Mining Example:** Email providers use anomaly detection on text patterns to detect spam or phishing emails that deviate from normal message content.

Experiment

Outlier Detection Using Isolation Forest and Z-Score with Real-World Case Study Analysis

Problem:

1. Use Isolation Forest or Z-score to detect outliers in real datasets.

2. Select a case study and perform end-to-end mining.

3. Document insights and business value.

Solution:

Step 1: Outlier Detection Using Isolation Forest on a Real Dataset

I’ve used the Credit Card Fraud Detection dataset (or any available real-world dataset with numeric features) to detect anomalies.

Python Code

```
import pandas as pd
from sklearn.ensemble import IsolationForest
import matplotlib.pyplot as plt

# Load dataset (example: credit card transactions)
data = pd.read_csv('creditcard.csv')

# Select numeric features for outlier detection
X = data.drop(columns=['Class'])  # Assuming 'Class' is label

# Initialize Isolation Forest
iso_forest = IsolationForest(contamination=0.01, random_state=42)
outliers = iso_forest.fit_predict(X)

# Add outlier flag to dataframe
data['outlier'] = outliers
data['outlier'] = data['outlier'].map({1: 0, -1: 1})  # 1 = outlier, 0 = normal

# Number of detected outliers
print(f"Detected outliers: {data['outlier'].sum()}")

# Visualize outliers on first two principal components (optional)
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

plt.scatter(X_pca[:, 0], X_pca[:, 1], c=data['outlier'], cmap='coolwarm', s=10)
plt.title('Isolation Forest Outlier Detection')
plt.xlabel('PCA Component 1')
```

```
plt.ylabel('PCA Component 2')
plt.show()
```

Step 2: Case Study – Market Basket Analysis in Retail

Scenario:
A retail company aims to enhance product placement strategies through insights gained from association rule mining.

- By analyzing transaction data, the retailer identifies frequent itemsets and generates association rules (e.g., *“Customers who purchase bread are also likely to purchase butter”*).
- These insights are presented on interactive dashboards (comparable to Tableau use cases) to inform decisions on shelf layout and promotional strategies.
- The outcomes include better cross-selling opportunities, enhanced customer experience, and increased overall sales.

Step 3: Documented Insights and Business Value

- **Fraud Detection:**
  Outlier detection techniques are used to flag suspicious transactions early, helping prevent financial losses that could reach millions.
- **Market Basket Analysis:**
  Enables personalized marketing campaigns and smarter inventory management, resulting in increased revenue and improved customer satisfaction.
- **Real-Time Data Warehousing:**
  Companies like Beyerdynamic utilize real-time data warehouses to quickly adapt to shifts in demand. This reduces excess inventory and enhances the accuracy of sales forecasting.
- **Transportation Optimization:**
  Data mining empowers ride-sharing companies like Uber and Lyft to refine route planning, lower fuel consumption, and decrease customer wait times—leading to significant improvements in operational efficiency.

Conclusion

In this experiment, we explored the practical application of outlier detection techniques such as Z-score and Isolation Forest on real-world datasets. These methods proved effective in identifying anomalies that could signify critical issues such as fraud, system failure, or unusual user behavior. By detecting these outliers early, organizations can take preventive actions, ensuring improved efficiency and risk management.

The selected case study demonstrated a complete data mining workflow — from data cleaning and anomaly detection to interpreting insights and linking them to real business value. For example, identifying unusual spending patterns in financial data can help prevent fraud, while spotting performance outliers in educational data can enable timely student intervention.

Overall, this experiment emphasized the importance of outlier detection not only as a technical tool but also as a strategic asset in decision-making across various industries. It also showcased how data mining adds value by uncovering hidden patterns and supporting more informed, proactive actions.