Chromatography Anomaly Detection System

Using One-Class SVM and LSTM with Streamlit UI

May 9, 2025

Introduction

Purpose: Detect anomalies in High-Performance Liquid Chromatography (HPLC) systems to ensure reliability and performance.

Why it matters: Anomalies (e.g., column clogging, high retention time) can lead to inaccurate results and system downtime.

Solution: A machine learning system combining One-Class SVM, LSTM, and a Streamlit dashboard for real-time monitoring.

Problem Statement

Challenges in HPLC:

- Complex data with multiple parameters (e.g., peak width, retention time).
- Anomalies are rare but critical (e.g., column overuse, contamination).
- Need for predictive maintenance to prevent failures.

Goal: Automatically identify historical and future anomalies, provide actionable insights, and visualize results intuitively.

Solution Overview

Key Components

One-Class SVM: Detects anomalies in historical data by learning normal behavior.

LSTM: Predicts future parameter values and potential anomalies.

GPT-4: Analyzes anomalies and suggests chromatography-specific causes (e.g., "column clogging").

Streamlit UI: Interactive dashboard for data visualization and anomaly monitoring.

Input/Output

Input: `data.csv` (HPLC data).

Output: `train_model/predictions.csv`, anomaly plots, and GPT-4 summaries.

Data Processing

Input Data: `data.csv` with columns like `injection_time`, `column_serial_number`, `peak_width_5`, etc.

Preprocessing Steps:

- Convert `injection_time` to datetime, remove timezone.
- Handle missing values using median imputation per column.
- Clip outliers at 99th percentile.
- Encode categorical columns (e.g., `system_name`) using LabelEncoder.
- Add features: `injection_count`, `days_since_start`.

Output: Cleaned DataFrame ready for model training.

Model Workflow

Historical Anomaly Detection

- **One-Class SVM**:
  - Trains on normal data (`nu=0.1, kernel='rbf'`).
  - Flags outliers as anomalies (`anomaly_flag`).
  - Computes `anomaly_score` and `anomaly_deviation`.

Future Predictions

- **LSTM**:
  - Uses sequences (`SEQ_LENGTH=10`) to predict 14 days ahead.
  - Bidirectional LSTM with 150 units, Huber loss.
- **Anomaly Detection**: Applies One-Class SVM on predictions.

Cause Analysis

- GPT-4 generates causes (e.g., "High peak_width_5 due to column clogging; recommend replacement").

Streamlit UI

Dashboard Overview

Purpose: Visualize historical and predicted anomalies, filter data, and summarize findings.

Access: Run `streamlit run app.py`, view at `http://localhost:8501`.

Key Features

Sidebar Filters: System Name, Method Set, Column Serial Number, Performance Metric.

Plots: Scatter plot of parameters with anomalies (historical: orange X, predicted: red X).

Deviation Graph: Shows anomaly deviations over time (1 day, 1 week, 1 month).

Tables: Historical and predicted data with anomaly details.

GPT-4 Summary: Tabular summary of future anomalies.

[Placeholder: Insert screenshot of Streamlit dashboard here]