Firefly Algorithm for Anomaly Detection for Red Wine Quality Dataset

Author: Jude Bariana, Ria Kokate
Course: Data Science
Date: 9th December 2025

Project Overview

This project focuses on detecting anomalies in red wine quality data using the Firefly Algorithm (FA). Anomalies are defined as wines with extremely low or high quality ratings. FA is a nature-inspired metaheuristic optimization algorithm based on the flashing behavior of fireflies. The main goal is to assign feature weights that maximize the detection of anomalous wine samples.

The report will include:

1. Abstract
2. Introduction
3. Dataset Description and Preprocessing
4. Methodology (Firefly Algorithm + Weighted Mahalanobis Distance)
5. Implementation Details
6. Results and Analysis
7. Mathematical Justification of Precision & Recall
8. Discussion and Conclusion
9. References

Abstract

This study presents an anomaly detection approach on the UCI Red Wine Quality dataset using the Firefly Algorithm (FA) combined with a weighted Mahalanobis distance metric. The objective is to identify wines with extremely high or low quality scores (considered anomalies) through an unsupervised, optimization-driven process. The FA optimizes feature weights to enhance anomaly separability by modeling the interaction of fireflies in a population-based search. Experimental results show effective performance in detecting extreme samples, with recall and precision behavior consistent with theoretical expectations for highly imbalanced datasets.

1. Introduction

Anomaly detection identifies rare items that deviate from normal patterns. In this project, anomalies correspond to wines of unusually low or high quality, as these samples may result from measurement errors or exceptional production conditions.

Traditional anomaly detection methods rely on distance metrics or clustering algorithms that treat all features equally. However, in correlated data such as physicochemical measurements of wine, certain features contribute more strongly to quality extremes. To capture this, we integrate a feature-weighted Mahalanobis distance optimized by the Firefly Algorithm, a nature-inspired metaheuristic proposed by Yang (2009).

The Firefly Algorithm simulates the social flashing behavior of fireflies, where brighter individuals attract others. In optimization terms, brightness corresponds to the quality of a candidate solution (here, the feature weights). This work aims to:

1. Identify anomalies using weighted multivariate distance.
2. Optimize feature relevance using FA.
3. Evaluate model performance via precision, recall, and F1 metrics.

2. Dataset Description and Preprocessing

The dataset was obtained from the UCI Machine Learning Repository:

Wine Quality — Red Wine (Cortez et al., 2009)

It consists of **1599** samples and **12** physicochemical features including: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality.

Anomaly Labeling

Samples with:

- quality ≤ 3 or quality ≥ 8 were labeled as anomalies (1)

- others labeled as normal (0)

This yielded **28 anomalies (1.75%) and 1571 normal samples (98.25%)**, creating a highly imbalanced dataset.

Preprocessing Steps:

- Loaded and cleaned CSV data.
- Converted all attributes to numeric values.
- Removed empty rows/columns.
- Standardized all feature columns using StandardScaler to ensure uniform scale.

3. Methodology

3.1 Weighted Mahalanobis Distance

The Mahalanobis distance measures how far a point lies from the center of a multivariate distribution:

$$D_M = [(x-\mu)^T \Sigma^{-1} (x-\mu)] \text{ sqrt}$$

where

- x is a feature vector,
- μ is the mean of normal samples,
- Σ is the covariance matrix of normal samples.

To assign varying importance to features, we define a weighted Mahalanobis distance:

$$D_W(x) = [(x-\mu)^T W \Sigma^{-1} W (x-\mu)] \text{ sqrt}$$

where,
W=diag(w1,w2,...,wd) is a diagonal matrix of feature weights.
These weights $w_i \in [0.05,1.0]$ are optimized by the Firefly Algorithm.

Higher w_i values amplify the feature's contribution to anomaly detection.
Lower w_i values suppress irrelevant or noisy features.

3.2 Firefly Algorithm

The Firefly Algorithm (FA) is based on three principles (Yang, 2009):

1. All fireflies are unisex — any firefly can be attracted to others.
2. Attractiveness is proportional to brightness and decreases with distance.
3. The brightness of a firefly is determined by the objective function.

Brightness (fitness): Each firefly represents a weight vector **w=[w1,...,wd]**
Its brightness corresponds to the fitness, defined as:

$$F(w)=AUC(w)-\lambda \ // \ w \ //_1$$

where

- AUC(w) is the area under the ROC curve for the weighted distance,
- λ is a sparsity penalty (set to 0.05),
- $// w //_1$ encourages simpler, interpretable weight vectors.

Movement equation:

$$w_i^{t+1}=w_i^t+\beta_0 e^{-\gamma r_{ij}^2}(w_j^t-w_i^t)+\alpha\epsilon$$

where

- β_0=1.0 = attractiveness at zero distance
- γ=0.6 = light absorption coefficient
- r_{ij} = Euclidean distance between fireflies i and j
- α=0.65 (decays by 0.997 per generation)
- ϵ = random perturbation vector

Algorithm settings:

Parameter	Value
Fireflies	80
Generations	60
Sparsity penalty	0.05
Cross-validation folds	5

4. Implementation and Training

1. Initialize fireflies with random weights in [0.05, 1.0].
2. Compute fitness of each firefly using cross-validated AUC of the weighted Mahalanobis distance.
3. Move less bright fireflies toward brighter ones using the movement equation.
4. Decay randomness each generation for fine-tuning convergence.
5. Select the best weight vector after 60 generations.
6. Use best weights to compute anomaly scores and threshold them using 70–98th percentile cutoffs.

FA successfully converged to a stable set of weights emphasizing features like **volatile acidity, alcohol, and density.**

5. Results and Analysis

5.1 Feature Importance

The optimized feature weights (w_i) highlighted:

- High weights: volatile acidity, density, alcohol
- Low weights: residual sugar, citric acid, pH

These align with chemical intuition — wines with abnormal acidity or alcohol content often deviate in quality.

5.2 Performance Metrics

Metric	Value (Example Run)
Precision	0.68
Recall	0.71
F1-score	0.69
ROC-AUC	0.93
PR-AUC	0.79

6. Mathematical Justification

Let:

- TP = true positives (correctly detected anomalies)
- FP = false positives (normal wines labeled anomalies)
- FN = false negatives (missed anomalies)

Then:

Precision=TP/TP+FP,

Recall=TP/TP+FN

Given **28** anomalies in total, even small misclassifications strongly affect the metrics.

For instance, if the model predicts: TP=20, FP=9, FN=8

Precision=20/29=0.69,
Recall=20/28=0.71

Mathematically, recall reflects sensitivity to missed anomalies (FN), while precision penalizes false alarms (FP) — both constrained by data imbalance.

7. Discussion

The Firefly Algorithm efficiently discovered optimal feature weights for anomaly detection:

- Adaptive weighting improved separation of normal vs. extreme wines.
- Weighted Mahalanobis distance leveraged feature correlations.
- Despite imbalanced data, FA achieved balanced precision–recall performance.

Observations:

- Increasing generations beyond 60 improved convergence stability.
- Moderate α-decay prevented premature stagnation.
- Using 5-fold cross-validation reduced overfitting.

Limitations:

- Computationally expensive ($O(n^2 \times \text{generations})$).
- Performance sensitive to hyperparameters.

8. Conclusion

This project demonstrates the integration of the Firefly Algorithm with weighted Mahalanobis distance for robust anomaly detection in wine quality data. The method effectively emphasized relevant chemical attributes, achieved high AUC values, and maintained interpretable feature contributions. Future work can include hybridization with other metaheuristics (e.g., Particle Swarm Optimization) or unsupervised extensions for unlabeled anomaly detection.

9. References

1. Yang, X. S. (2009). *Firefly Algorithms for Multimodal Optimization.* Stochastic Algorithms: Foundations and Applications. Springer.

2. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Modeling Wine Preferences by Data Mining from Physicochemical Properties.* Decision Support Systems, 47(4), 547–553.

3. Chandola, V., Banerjee, A., & Kumar, V. (2009). *Anomaly Detection: A Survey.* ACM Computing Surveys, 41(3), 1–58.

4. Nayak, J., Naik, B., & Behera, H. S. (2020). *A Systematic Review on Firefly Algorithm: Past, Present and Future. Artificial Intelligence Review, 53(2), 1027–1077.*

5. Sharma, V., & Gupta, S. (2023). *An Enhanced Outlier Detection Approach for Multidimensional Datasets Using a Synergistic Firefly and Grey Wolf Optimization-Based Method. Instrumentation Mesure Métrologie, 28(3), 425–434.*