Tesina 2

Analisi Algoritmo Genetico come Feature Selection su Dataset DARWIN

Contesto del Problema

Un team di ricerca deve analizzare l'impatto dei diversi parametri degli Algoritmi Genetici (GA) sulla loro performance e convergenza utilizzando il dataset DARWIN. Questo dataset contiene caratteristiche di scrittura a mano per lo studio dell'Alzheimer, offrendo un contesto reale e significativo per l'analisi degli algoritmi genetici.

Specifiche del Dataset

- DARWIN Dataset:
  - Features di scrittura a mano
  - Caratteristiche cinematiche
  - Pressione della penna
  - Parametri geometrici
  - Caratteristiche temporali
- Complessità:
  - Multiple feature categories
  - Dati numerici continui
  - Correlazioni complesse

Obiettivi

1. Implementare un **Algoritmo Genetico** base per feature selection
2. Studiare sistematicamente l'effetto dei parametri:
   - Convergenza dell'algoritmo
   - Stabilità delle soluzioni
   - Velocità di esecuzione

- Robustezza della selezione

3. Determinare configurazioni ottimali per:
    - Diverse dimensioni del subset di feature
    - Vincoli computazionali
    - Requisiti di stabilità

Vincoli

- Utilizzo stesso seed per confronti equi
- Minimo 30 run per configurazione
- Tempo massimo di esecuzione per run
- Gestione appropriata missing values

Fasi del Progetto

Fase 1: Implementazione GA Base

- Sviluppare algoritmo genetico con:
    - Codifica binaria per selezione feature
    - Funzione fitness basata su correlation analysis
    - Operatori genetici modulari
- Implementare logging dettagliato:
    - Fitness per generazione
    - Feature selezionate
    - Tempi di esecuzione
    - Diversità popolazione

Fase 2: Analisi Parametrica

1. **Scenario Dimensione Popolazione**
    - Test dimensioni: [20, 50, 100, 200, 500]
    - Metriche:
        - Velocità convergenza

        - Stabilità selezione feature
        - Costo computazionale
    - Altri parametri fissi:
        - Crossover: 0.8
        - Mutazione: 0.1
2. **Scenario Operatori Genetici**
    - Crossover rates: [0.6, 0.7, 0.8, 0.9]
    - Mutation rates: [0.01, 0.05, 0.1, 0.15]
    - Metodi selezione:
        - Tournament (k=2,3,4)
        - Roulette Wheel
    - Popolazione fissa: 100
3. **Scenario Criteri di Stop**
    - Numero generazioni fisse: [50,100,200]
    - Convergenza (no improvement):
        - Soglie: [10,20,30] generazioni
        - Tolleranze: [1e-4, 1e-5, 1e-6]

Output Richiesti per Ogni Scenario

- Curve di convergenza
- Box plot distribuzioni fitness
- Frequenza selezione feature
- Tempi di esecuzione