FAST - NUCES

Assignment - 02

Genetics Algorithm

Bilal Akbar - 21i-1174
Ibraheem Rehman - 21i-1102

Artificial Intelligence

Mam Shahela Saif

March 30, 2024

Data Extraction

Purpose

The data extraction process aims to filter and extract specific information from the RAVDESS Facial Landmark Tracking dataset, focusing on speech modalities and emotions such as happiness and sadness.

Steps

1. ### Data Acquisition
   - Obtain the RAVDESS Facial Landmark Tracking dataset from the provided sources, either from Kaggle or Zenodo.
2. ### Filtering Relevant Entries
   - Extract data entries corresponding to speech modalities (modality code 03) and emotions happy (emotion code 03) and sad (emotion code 04) as per the assignment requirements.
3. ### Dataset Preparation
   - It initializes two empty DataFrames: happy_data and sad_data.
4. ### Iterate Through CSV Files
   - For each CSV file in the specified directory:
     - Extract relevant information from the filename (e.g., modality, voice, emotion, etc.).
     - Check if the file matches the desired criteria.
     - Read the CSV data into a DataFrame.
     - Drop unwanted columns.
5. ### Filter and Append Data
   - Based on the emotion code:
     - If emotion is “03” (happy), append data to happy_data.
     - If emotion is “04” (sad), append data to sad_data.
6. ### Print Summary
   - Total number of rows in happy_data and sad_data.
   - Total number of columns in the data.

Solution

Purpose

The purpose of this script is to implement a genetic algorithm (GA) for feature selection to enhance the performance of a neural network model in classifying emotions. By selecting the most relevant features from the dataset, the GA aims to optimize the neural network's ability to accurately classify emotions based on facial landmarks.

Genetic Algorithm

The genetic algorithm utilizes a binary representation for chromosomes, with each gene representing the presence or absence of a specific facial landmark feature. By iteratively evolving a population of chromosomes, the algorithm seeks to identify the optimal subset of features that maximize classification accuracy.

Key Components

1. **Population Initialization (Chromosome Representation)**
   - Each chromosome represents a subset of features selected for the neural network model.
     - The chromosome is represented as a binary string, where each gene corresponds to a feature.
     - If the gene is set to 1, the corresponding feature is selected; otherwise, it is not selected.
     - The length of the chromosome is equal to the total number of features in the dataset.
2. **Fitness Function**
   - A fitness function evaluates the quality of each chromosome by training a neural network model on the selected features and measuring its accuracy.
   - The accuracy serves as the fitness score, indicating how well the neural network performs with the given subset of features.

3. **Genetic Operators**
   a. **Selection Mechanism**
      - The algorithm employs a roulette wheel selection method to select parent chromosomes based on their fitness scores.
      - Chromosomes with higher fitness scores are more likely to be selected as parents for reproduction.

   b. **Crossover**
      - Single-point crossover is applied to pairs of parent chromosomes to generate offspring.
      - A random crossover point is chosen, and the genes beyond that point are exchanged between the parents to create new combinations of features.

c. Mutation

- Bit-flip mutation is performed on the offspring chromosomes with a certain probability.
  - Each gene in the chromosome has a certain probability of being flipped from 0 to 1 or from 1 to 0.
- This introduces small random changes in the chromosome's genes, promoting genetic diversity within the population.

Steps

1. Initialization

- Following genetics related variables are initialized that are involved in the genetics algorithm working:

```
population_size = 10
parent_crossovers = population_size // 2
total_generations = 2 #The number of generations
mutation_probability = 0.1
total_genes = 0 #These are set based on the no. of columns in the data file
```

- Initialize the population with random chromosomes.
- Print the initial population and the fitness of the first chromosome

2. Generation Loop

- Iterate through a predefined number of generations.
- Calculate the fitness for each individual in the population.
- Calculate the total fitness of the population.

3. Selection and Reproduction

- For each generation, create a new population.
- For each individual in the new population:
  - Select two parent chromosomes using the roulette wheel selection method based on their fitness.
  - Generate a child chromosome through crossover of the selected parents.
  - Optionally, apply mutation to the child chromosome based on a predefined mutation probability.

4. Evaluation and Printing

- Print the details of each generation, including the accuracy of each chromosome.
- If it's the final generation, print the accuracy of each chromosome and identify the best chromosome and its accuracy

- Result of the Final Model:

```
population[0] fitness: 0.7994
population[0] fitness: 0.7994
Inital Generation:
        Chromosome 1 accuracy: 0.7994
        Chromosome 2 accuracy: 0.6999
        Chromosome 3 accuracy: 0.8053
        Chromosome 4 accuracy: 0.8375
        Chromosome 5 accuracy: 0.7862
        Chromosome 6 accuracy: 0.8565
        Chromosome 7 accuracy: 0.7423
        Chromosome 8 accuracy: 0.9341
        Chromosome 9 accuracy: 0.5593
        Chromosome 10 accuracy: 0.7877
Generation 2:
        Chromosome 1 accuracy: 0.7643
        Chromosome 2 accuracy: 0.5754
        Chromosome 3 accuracy: 0.8433
        Chromosome 4 accuracy: 0.8331
        Chromosome 5 accuracy: 0.6867
        Chromosome 6 accuracy: 0.5593
        Chromosome 7 accuracy: 0.7731
        Chromosome 8 accuracy: 0.6999
        Chromosome 9 accuracy: 0.6457
        Chromosome 10 accuracy: 0.8170
Generation 3:
        Chromosome 1 accuracy: 0.8360
        Chromosome 2 accuracy: 0.7994
        Chromosome 3 accuracy: 0.7599
        Chromosome 4 accuracy: 0.6808
        Chromosome 5 accuracy: 0.8463
        Chromosome 6 accuracy: 0.7848
        Chromosome 7 accuracy: 0.6750
        Chromosome 8 accuracy: 0.5754
        Chromosome 9 accuracy: 0.5593
        Chromosome 10 accuracy: 0.6486
```

```
Final Generation:
        Chromosome 1 accuracy: 0.8258
        Chromosome 2 accuracy: 0.7643
        Chromosome 3 accuracy: 0.6457
        Chromosome 4 accuracy: 0.8097
        Chromosome 5 accuracy: 0.6662
        Chromosome 6 accuracy: 0.8082
        Chromosome 7 accuracy: 0.8492
        Chromosome 8 accuracy: 0.7174
        Chromosome 9 accuracy: 0.6413
        Chromosome 10 accuracy: 0.7438

Resulting chromosome of the GA function:
Best chromosome: [1 1 1 1 1 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 0 1 1
 1 0 0 1 1 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1 1 0 1 1 1 1 1 1 0 1 0 1 0 1 0 1 1
 1 0 0 1 1 0 1 0 0 1 0 1 0 0 1 0 0 1 1 0 0 1 0 1 0 1 1 0 0 1 0 0 1 1 0 1 1
 1 0 1 0 1 1 1 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 0 1 1 0 0 1 1 0 1 0 1 1 1 1 0
 1 0 0 0 1 0 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 0 1 1 0 0 1 0 0
 1 1 0 0 1 1 0 1 1 1 1 1 0 0 1 1 1 0 0 0 0 0 1 1 0 1 0 1 1 0 1 1 1 0 0 0 1
 1 1 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 0 0 1 1 1 0 0 0 0 0 0 1
 1 0 0 1 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 1 1 1 0 1 1 0 1 1 0 0 1 0 1 1 1 1 0
 1]
Best accuracy: 0.8492

Time taken: 230.33 seconds
```

5. Termination Conditions

- After iterating through all generations, return the best chromosome and its accuracy.

6. Output

- Print the resulting best chromosome and its accuracy as the output of the genetic algorithm function.

Impact of Feature Selection on model performance and the features selected by the GA

- Feature selection is an important step in building a neural network model, as it can reduce the dimensionality of the data, improve the model's performance, and reduce overfitting.
- The GA implemented in Solution.py was used to select the most relevant features which are as follows:
    - Emotion Classification (happy and sad)
    - Intensity (normal)
    - Repetition (1st)
    - Actors (1 - 8)
    - Vocal Channel (speech)
        - With these features selected, the model was able to compute the best chromosomes in <u>3 Generations with a time of 185 seconds.</u>

```
Resulting chromosome of the GA function:
Best chromosome: [0 1 1 0 1 1 0 0 0 1 0 1 0 0 1 1 1 1 0 1 0 0 0 1 1 0 0 1 0 1 1 0 1 0 0 1 0
 0 0 1 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 0 0 0 1 1 0 0 0 1 1 0 1 1 1 1 0 0 1 0
 0 0 1 0 0 1 0 0 0 1 1 1 0 0 0 1 1 0 1 0 1 0 1 1 0 1 0 1 1 1 0 0 1 1 0 0 0
 1 0 1 0 1 1 1 0 1 0 1 1 0 0 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1 0 0 0 0
 1 0 0 1 0 1 0 0 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 1 1 1 1 1 0 0 0 1 1 0 1 0 1
 1 0 0 1 1 1 0 0 1 0 1 0 1 1 1 1 0 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1
 0 1 1 1 1 0 0 1 0 1 0 1 1 0 0 1 1 1 0 0 1 1 1 1 1 0 1 0 1 1 0 1 1 0 1 0 1
 0 0 0 1 0 0 0 1 1 0 0 1 0 1 0 1 1 1 1 1 1 1 0 0 1 0 1 0 0 0 1 1 1 1 1 1 0 0
 1]
Best accuracy: 0.9224
Time taken: 185.76 seconds
```

- However, if we increase the number of features to include:
    - All vocal channels
    - All statements
    - All actors
    - All repetitions
    - Emotion Classification (happy and sad)
    - Intensity (normal)
        - With these features selected, the model was able to compute the best chromosomes in <u>3 Generations with a time of 1077 seconds.</u>

- ```
  Resulting chromosome of the GA function:
  Best chromosome: [1 0 1 0 0 1 1 0 1 1 1 1 0 1 1 0 0 0 0 0 0 1 0 1 1 0 0 1 1 0 1 0 1 1 0 0 0
   1 0 1 1 0 0 1 0 0 0 0 1 0 1 1 1 1 0 1 1 0 0 1 0 1 0 1 0 1 1 0 0 1 0 1 0 1
   0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 1 1 1 1 1
   0 0 1 1 0 1 0 0 1 1 0 0 1 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 1 1 0 1 0 1 1 1 0
   0 0 0 0 1 0 1 1 1 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 1 0 1 0 1 0 0
   1 1 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0
   1 1 1 1 0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 1 0 0 1 1 0 0 0 1 1 1
   1 1 1 0 1 1 0 1 0 1 0 0 1 1 1 0 0 0 1 0 1 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 0
   0 1 1 0 1 0 0 1 0 0 1 1 1 1 1 1 0 1 1 0 0 1 0 0 0 1 0 1 1 1 1 1 1 1 0 1 1
   1 0 0 1 0 1 1 1 0 1 0 1 0 1 0 1 1 0 1 1 1 0 1 0 1 1 0 0 1 1 0 0 1 0 1 1 1
   0 1 1 0 0 0 1 1 0 0 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 0 0 1 1 0
   1 0 0 0 1 1 0 1 1 0 1 1 1 1 0 0 1 1 1 0 0 1 0 0 0 1 1 0 0 1 1 0 0 0 1 1 1
   1 1 0 1 1 1 1 1 0 0 0 0 1 0 1 1 1 1 0 0 1 0 0 0 0 0 1 1 1 1 1 0 0 0 1 1 0
   1 0 0 1 0 0 0 1 0 0 1 1 1 1 1 0 0 0 0 1 1 1 0 0 0 0 1 0 0 0 1 1 1 0 1 0 0
   1 1 0 0 0 1 0 1 0 1 0 1 1 1 1 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0 1 1 1 1 1 1
   0 0 1 0 0 1 1 1 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1
   0 1 1 1 1 0 0 0 1 1 0 1 1 0 1 1 1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 1
   1 1 0 0 0 0 1 1 1 0 1 1 1 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 0 0 1 0 1 1 1 0 0
   0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 1 0 1 0 0 0 1 0 1 1 0 0 0 0 1 1 0 0 1 1
   1 0 0 1 0 0 0 0 1]
  Best accuracy: 0.9455
  Time taken: 1077.26 seconds
  ```

- Although the time increases, it should be noted that the accuracy of the chromosome also increases.
  ```