CSCI4343 Project: Coffee Pour Over Human Feature vs. AI Challenge

CSCI4343 - Data Mining Final Project

November 2024

1 Introduction

In this project, you will try to become a coffee data miner. Pour over techniques is a very popular techniques that used in a lot of coffee shop. But to make a delicious coffee is not very simple. For example, to achieve the best coffee brewing performance, different world champion invented different type of strategies. If you like to know, here is a incomplete listed:

- 46 Method
- Rao Method
- Five pour Receipt
- Centre Pour Method

There are many different variations of filters and the design of pour over system, V60, Kalita, Origami, each one has its own features. A delicious cup of coffee often obtained by a very .

With the advance of sensor techniques, now, we could monitor the pour over method via **data**! We placed a smart scale to measure the water volume increasing throughout the process. You responsibility is: **Can we predict the origin of the coffee bean from the data?**. You will provide a set of data and the class label

2 Submission Guideline

Please complete the blackboard team member submission via the corresponding blackboard link and including your all team members. **Each team can have up to 4 students**.

The final project requires you to submit:

- a report (up to 3 pages) consists of introduction, methodology you used in the project, and experiment result

- Source code you used and the running result. If you use human crafted feature, please upload the crafted features as a csv file.
- Readme file about how to run the source code
- Feedback of this project

Additionally, you need to mention the responsibility of each student in the project report. **Google Colab** is highly recommended.

For each team we have:

- **Team with 1-2 Members**: You can choose to do *Task 2.1, Task 2.2, and Task 2.5* or *Task 2.3, Task 2.4, and Task 2.5*.
- **Team with 3-4 Members**: You need to do all Task 2 tasks.

3 Data Description

Like our Katydid vs. Grasshopper example in the class, we need to analyze the features. A preprocessing step is needed to obtain a set of useful features. You need to conduct the following steps:

An example of the data is shown in the following figure where x-axis is time stamp and y axis is the water volume.

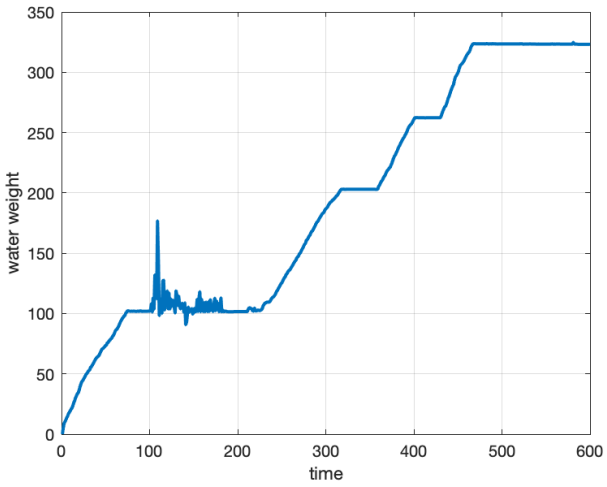

Figure 1: An Example of Brewing Data

The loading code is already provided in the corresponding google colab code. For each time series, it has the following format:

brew data: water volume collected over time. It is stored in an np array: "coffee_dataset".

origin of the coffee bean: For each brew data, we have a label labeling the original of the data. It will be either "Sprout" or "Rwanda". It is stored in "coffee_labels"

4 Task 1: Data Preprocessing

The goal of this project is "**build a data mining model such that, given the input of the brewing data, it guess the origin of this cup of coffee**.

In task 1, you need to complete the following pre-processing tasks:

Task 1.1: Visualize samples (5 pt) Loading data and use python (package **matplotlib**) to plot the data.

Task 1.2: By visualizing such data, answer which features you think is the best to form a data mining problem? Write down your answer in your submitted report.

Task 1.3: Convert the text label "Sprout" and "Rwanda" to numbers so that we could implement our classification model. To do so, you can form a new label array such that replacing all "Sprout" to 0 and "Rwanda" to 1. This array will be useful for Task 2

5 Task 2: Training Model (15 pt)

After you extracted the feature, next step, we will conduct a human feature vs. machine feature challenge. We have four steps in this part:

Task 2.1: Train a MLP model using all the data labeled data and output is two classes to perform classification.

Task 2.2: Taking cross validation into account, train a MLP model using 60% of labeled data and use the other 40% of labeled data to adjust which learning rate you should use and which epoch you should stop. Write down your observation in your submitted report.

Task 2.3: Train a MLP model to classify the data based on the features you summarized in Task 1.3 and output is two classes to perform classification.

Task 2.4: Similiar to Task 2.2, taking cross validation into account, train a MLP model using 60% of labeled data and use the other 40% of labeled data to adjust which learning rate you should use and which epoch you should stop. Write down your observation in your submitted report.

Task 2.5: I will provide 4 unlabeled coffee data in the later stage of the project. Please write down your model output.

bonus 2pt: team who get all 4 samples correct will recieve extra bonus.