1071 類神經網路在心理學上的應用 專題

Predict cuisine style of recipe based on ingredients

柳桓任 106753040, 朴在明 105702024

1. Introduction

Nowadays, food does not exist only as a meal to fill the stomach. Those who want to know healthy recipes with an emphasis on health, those who are curious about foreign foods and want to know how to cook in foreign countries, those who want to know how to cook gourmet food and make delicious food in a mixed way. People has various approach depend on owns reasons and topics.

Research on cuisine and recipe already study whether people prefer which flavor, which style of food they can make with a given ingredient, whether they can develop new recipes based on their preferences, and so on.

All dishes can have different results depending on which ingredients are used. Does the taste that people feel delicious are habitually domesticated by the food that eat when grown? Or is it a taste that feels delicious in common to all nations? If the latter is the case, why would the recipes and tastes change in the style of the country when the food of one country is spread to other countries?

We decided to analyze the locality of food, the most traditional approach to cooking as the first approach to it. The aim is to create a model that judges which cuisine belongs to a local style with given ingredients. Before learning the neural network learning model, the learning classification was done by the method of machine learning. In order to judge whether the effect of neural network is better than machine learning, we also conducted machine learning. We also confirmed whether data preprocessing used in machine learning can reduce errors in neural network learning models such as overfitting.

2. Dataset

We get dataset from one of kaggle competition named “whats cooking”, dataset has data from yummly.com which is online sharing recipe site. Format is json, train data include id of recipes, cuisines of recipes as label, ingredients which contents recipes. Test data include former ones except cuisine label.

In the overall cuisine classification of train data, the recipes belonging to Italy are about 8000, and the style that is the least, contains about 500. Due to the large number of western users, the number of ingredients frequently used in oriental recipes may be appear low.

3. Data analysis and data preprocessing

When we look at the top twenty ingredients(fig1.) and the top twenty ones of each style, the distribution of the ingredients is not similar. We found this ingredient as a special material. When a human being categorizes recipes contained this special ingredient, it is easy to distinguish which country food it is,

but when applied to a computer, over fitting can occur and special materials should be removed.

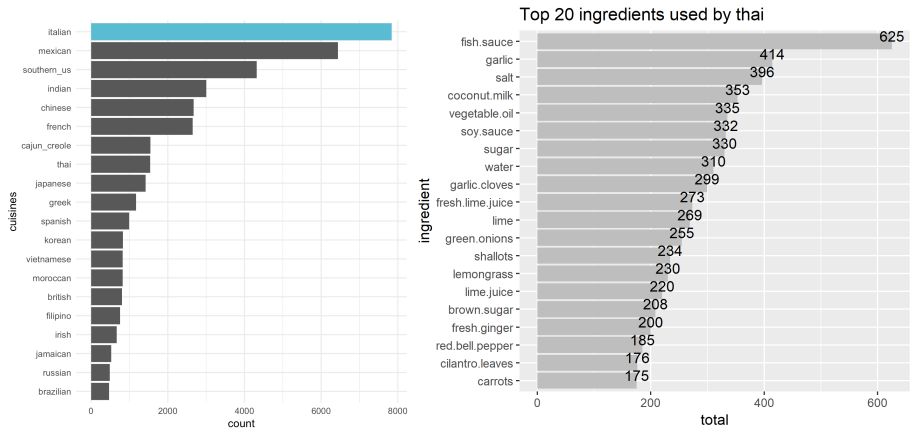

figure1. top 20 ingredients and top 20 ingredients used by thai

To prevent overfitting, I had to think about which ingredient should be removed considered as outliers. First type of outlier is the ingredient contained in too small in dataset, it was determined as the distribution of the entire data set.

Summary of Ingredients Used Times

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
1	1	4	64.15	20	18048

Second one is recipes contained large number of ingredients, it was determined as the total number of ingredients contained in the recipe, showing a long tail in the distribution(fig2.). Last one is whether or not the material appears as a whole style or a special cuisine(fig3.). Entropy was used to find out. These three parameters were used to process the input data as 5, 3.8, and 30.

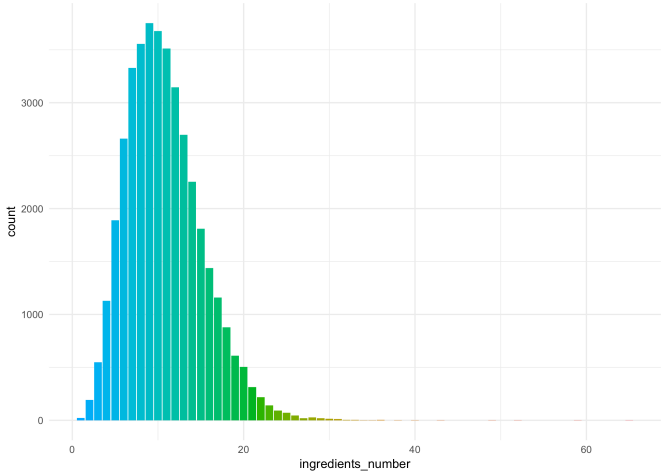

figure2. distribution of recipe

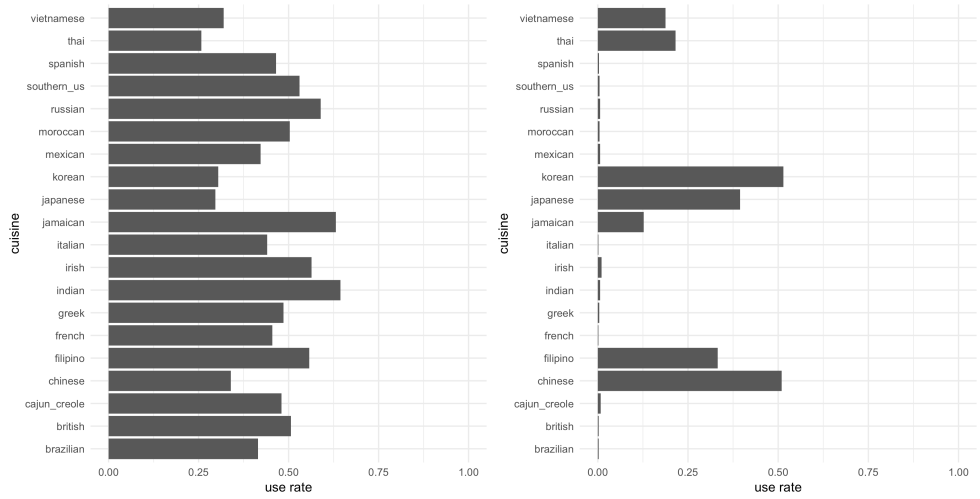

figure3 proportion of ingredient ; salt(left) , soy sauce(right)

4. Modeling and results

We make neural network model using keras library in r. and make machine learning model as comparion.

The first try with machine learning models(random forest, decision tree, k-NN, naïve bayes) get 0.2 as accuracy, we only do simple data cleaning in this try. We find this model classified all recipes to Italian, it seems null model because of proportion of data by cuisine. Second try with machine learning models get 0.68 as accuracy, we do data cleaning and third one of data preprocessing. We find the power of data preprocessing in machine learning. Third try get 0.73 as accuracy, with removing outlier.

It is big difference in machine learning and neural network learning in this study, in machine learning the data preprocessing is important to arise accuracy, oppositely in neural network the data preprocessing has not much effects, even if has negative effects like our results. We can check it NN with data preprocessing.R has 0.78 accuracy despite of NN with data preprocessing.R has 0.79 accuraacy.

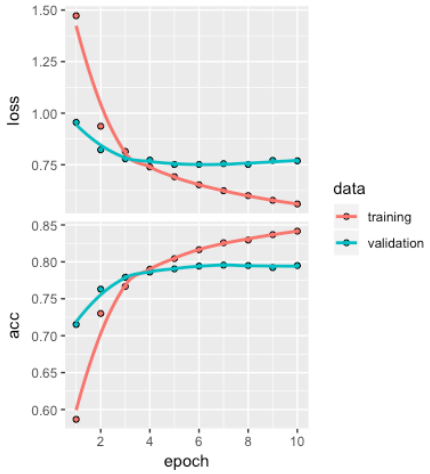

figure4. train model accuracy in NN