Use recipe ingredients to categorize the cuisine

Manshan Lin, Xin Tong

Different countries and regions have different cooking habits and corresponding recipes, and it seems it’s hard for us to determine the cuisine of a recipe based on its ingredients. When people happen to choose an unfamiliar restaurant, it’s quite difficult to expect what kind of food they will get. Accordingly, training a classifier to predict the recipe based on its ingredients seems useful and interesting.

There are two reasons to choose this task. First, feature engineering plays an important role in this task which can cultivate our ability to preprocess the data with several useful techniques. To be more specific, there are several thousand kinds of ingredients while there are only twenty cuisine categories. Obviously, directly using KNN or decision tree may encounter dimensions’ explosion. Decreasing the dimension and choosing effective features is the key point for us. What’s more, this model is generalizable which can be extended to solve other problems. For example, it’s similar to classify user characters with their pet phrases. Many classification problems with high dimension, uncertain input, and few output can also be solved by this classifier.

We obtain the original data from. They provide nearly 40000 training samples and 10000 test samples. We will train classifier and submit our test result to get corresponding accuracy from Kaggle. Each training sample comes with a unique id, a “cuisine” label referring its category and several ingredients. Basically, cuisine is our training target and ingredients are our inputs.

We are planning to do this project in following steps.

* ● Feature engineering. We will use Bag of Words to build an ingredient dictionary and  represent each recipe with a word vector of ingredient occurrence in ingredients dictionary. We may also adopt PCA to avoid dimension explosion, since there are over 6700 ingredients and directly using data will make decision tree too deep.
* ● Pre-processing and build classifier. We could use KNN, random forest and other classification algorithms to handle data and get an overview without cleaning data. For example, we could apply “and” operation on different recipe word vector pairs to represent the similarity between them then we use KNN to classify cuisine. We will evaluate our classifier and clean data based on results, since there are some ingredients can be seen as parents and children. (e.g. “grape tomato” and “tomato”)
* ● Re-train. After cleaning data and finishing PCA, we will try logistic regression or Bayes classification in lower dimension and compare performance with previous classifiers.
* ● Ensemble classifier. We believe different classifiers have different advantages over this dataset. We hope to combine all of advantages so assembling algorithm will be used to get a better classifier.